Strategic Communication, Voting and Political Institutions: Essays in Behavioral Political Economy

Dissertation

zur Erlangung des akademischen Grades
doctor rerum politicarum
(Dr. rer. pol.)

vorgelegt dem Rat der Wirtschaftswissenschaftliche Fakultät
der Friedrich-Schiller-Universität Jena
am 02.11.2016

von M.Sc. Isabel Marcin

geboren am: 26.06.1986 in: Würzburg, Deutschland

Gutachter

Prof. Dr. Oliver Kirchkamp, Friedrich-Schiller-Universität Jena

Prof. Dr. Christoph Engel, Max-Planck-Institut zur Erforschung von Gemeinschaftsgütern, Bonn

Datum der Verteidigung: 31.01.2017

Acknowledgements

Writing this thesis would have been impossible without many wonderful people to whom I would like to express my deep gratitude.

First of all, I would like to thank Christoph Engel for accepting me as a PhD student at the Max Planck Institute for Research on Collective Goods and the International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World. I have tremendously benefited from the research environment at the institute where I spent three and a half wonderful years.

I am particularly grateful to both my supervisors, Christoph Engel and Oliver Kirchkamp, for their invaluable guidance, support, and comments during the entire work on this thesis.

I owe special thanks to my co-authors Mark Le Quement, Pedro Robalo and Franziska Tausch for many hours of stimulating discussions and great collaborations.

For helpful discussions and comments on earlier versions of the dissertation I would like to thank my MPI colleagues Benjamin Bachi, Susann Fielder, Adrian Hillenbrand, Michael Kurschilgen, Monika Leszczyńska, Nicolas Roux as well as Guillaume Frechette, Sebastian Goerg, Jens Grosser, Sebastian Kube, Dimitri Landa, Rebecca Morton, and Andrew Schotter.

Finally I would like to thank Michael, my family, and my friends for their love and support.

Deutsche Zusammenfassung

Fast alle politischen Institutionen haben das Ziel Probleme kollektiver Entscheidungen und kollektiven Handelns zu lösen. Häufig wird das zugrundeliegende Problem durch fehlendes Wissen über die Präferenzen anderer Individuen oder durch unvollständige Information verursacht. In einem sozialen Dilemma würden zum Beispiel viele Individuen kooperieren, wenn sie wüssten, dass auch andere dazu bereit wären (Fischbacher und Gaechter, 2010). In Komitees, wie sie Condorcet (1785) in seinem Condorcet-Jury-Theorem betrachtet, würden die Mitglieder alle die gleiche Entscheidung favorisieren, wenn sie die gleiche Information besäßen. Wenn sie jedoch über private, potentiell voneinander abweichende Informationen verfügen, ist es möglich, dass sie in ihren präferierten Entscheidungen nicht übereinstimmen.

Diese Dissertation analysiert die Rolle von Kommunikation und Abstimmungen in Gruppen, in denen Interessenkonflikte vorhanden sind, und untersucht die daraus resultierenden Konsequenzen für das Design von Institutionen. Der Schwerpunkt dieser Arbeit liegt auf zwei Themen. Kapitel 2 und 3 studieren das Potential von Kommunikation, Informationen zu aggregieren. Insbesondere wird untersucht, ob Nicht-Standard-Präferenzen Individuen veranlassen, in strategischen Situationen die Wahrheit zu sagen. Kapitel 4 untersucht die Auswirkungen eines demokratischen Entscheidungsverfahrens auf Sanktionierung und Regelkonformität.

Diese Dissertation trägt zur jungen Disziplin "Behavioral Political Economy" (verhaltensorientierte Politische Ökonomie) bei, die im Gegensatz zum standardökonomischen Ansatz, der auf starken Rationalitätsannahmen beruht, das tatsächliche Verhalten von Menschen in politischen Institutionen untersucht (Fischbacher, Hillman und Ursprung, 2015; Schnellenbach und Schubert, 2015). Diese Disziplin folgt damit sowohl dem Ansatz der Verhaltensökonomie, deren Ziel es ist, Standardmodelle durch Theorieerweiterungen realistischer zu gestalten, als auch der experimentellen Ökonomie, die Wert auf die Überprüfung theoretischer Modelle legt (Camerer, 2003).

In drei eigenständigen Kapiteln verwende ich theoriegeleitete Laborexperimente und integriere verhaltensökonomische Motive in Standardmodelle: kognitive Hierarchie und soziale Präferenzen in Kapitel 2, soziales Image in Kapitel 3 und Legitimität in Kapitel 4. Während standardökonomische Modelle der Politischen Ökonomie mit ihren Annahmen von eigennützigen Präferenzen und voller Rationalität wesentliche Einblicke in politische Prozesse geleistet

haben (Rowley, 1993), können sie andere Phänomene nicht erklären. Zu diesen Phänomenen gehören Überkommunikation und Regelkonformität, die im Folgenden näher erläutert werden. Ich stelle Erklärungen für beide Phänomene auf und teste diese experimentell.

Condorcet (1785) hat die Idee eingeführt, dass politische Institutionen private Information aggregieren können. Abstimmungen werden als ein Mittel der Kommunikationen betrachtet: Mit der Abstimmungsentscheidung offenbaren Wähler ihre private Information durch sogenanntes ehrliches Abstimmen.[1] Das Condorcet-Jury-Theorem zeigt auf, unter welchen Bedingungen die Mehrheitsregel effiziente Informationsaggregation durch ehrliches Abstimmen ermöglicht; spätere Erweiterungen offenbaren die Grenzen dieses grundsätzlichen Ergebnisses (Austen-Smith und Banks, 1996; Feddersen und Pesendorfer, 1998; Piketty, 1999). So führen zum Beispiel heterogene Präferenzen innerhalb von Gruppen zum Zusammenbruch des ehrlichen Abstimmungsverhaltens (Coughlan, 2000). In diesen Situationen können andere Institutionen, wie etwa Debatten, komplementär zu Abstimmungen wirken und einen wirksamen Informationsaustausch ermöglichen.

Während Kommunikation einer direkteren Form der Informationsaggregation entspricht, kann auch diese scheitern (Coughlan, 2000). Dies geschieht vor allem in Situationen, in denen die Präferenzen der Akteure deutlich voneinander abweichen und sie unabhängig von der Informationslage unterschiedliche Politiken bevorzugen. Ein bedeutendes Beispiel ist die Lobbyarbeit, in der informierte und auf ihren eigenen Vorteil bedachte Lobbyisten Nachrichten an uninformierte, aber befugte Entscheidungsträger senden (Crawford und Sobel, 1982). Obwohl die Standardtheorie hier strategische Kommunikation prognostiziert, finden viele Experimente, dass Individuen selten Nachrichten versenden, die nicht wahrheitsgetreu sind (für einen Überblick siehe Abeler, Nosenzo und Raymond, 2016). Dieses Phänomen wird in der Literatur Überkommunikation genannt.

In Kapitel 2 und 3 werden verschiedene Formen der Überkommunikation untersucht. Im letzten Jahrzehnt hat die Literatur dafür zahlreiche verhaltensökonomische Erklärungen vorgebracht. Diese beruhen entweder auf Präferenzen (z.B. eine Präferenz für Ehrlichkeit oder eine Präferenz als eine ehrliche Person wahrgenommen zu werden) (z.B. Ellingsen und Östling, 2010; Fischbacher und Föllmi-Heusi, 2013; Kartik, Ottaviani und Squintani, 2007; Mazar,

[1] Im englischen heißt der Fachbegriff sincere voting. Der Wähler stimmt für diejenige Alternative, die er am meisten bevorzugt, ohne strategische Interaktionen zu berücksichtigen.

Amir und Ariely, 2008; Sánchez-Pagés und Vorsatz, 2007), einer sozialen Norm (d. h. Individuen wollen sich konform der Norm, die Wahrheit zu sagen, verhalten)(Diekmann, Przepiorka und Rauhut, 2015; Rauhut, 2013) oder begrenzter Rationalität (Cai und Wang, 2006). Einige Motive sind je nach Kontext wichtiger als andere. So können bei kollektiven Entscheidungen, die viele individuelle Entscheidungen aggregieren, soziale Präferenzen eine bedeutende Rolle spielen. In Kapitel 2 untersuche ich gemeinsam mit Mark Le Quement, ob Mitglieder eines Komitees anstatt der Maximierung ihres eigenen Profits das Ziel haben, die Summe aller individuellen Profite zu maximieren. Dies sollte zur Folge haben, dass sie in der Deliberationsphase vor der Abstimmung ihre private Information wahrheitsgemäß teilen. Angesichts der komplexen Entscheidungssituation überprüfen wir außerdem, ob die Wahrheit zu sagen als eine Heuristik verwendet wird.

Kapitel 2 analysiert also insbesondere Situationen, in denen Mitglieder eines Komitees im Konflikt zwischen dem sozialen Ziel, eine korrekte Entscheidung zu treffen, und ihrem individuellen Ziel der Profitmaximierung stehen. In Condorcet Jurys, in denen Spieler über private Informationen verfügen, kann der wahrheitsgemäße Austausch von Informationen in einer Deliberationsphase die Qualität der Gruppenentscheidung verbessern, d. h. die Wahrscheinlichkeit erhöhen, dass diese richtig ausfällt (z.B. einen Angeklagten frei sprechen, der unschuldig ist). In Gruppen mit heterogenen Präferenzen haben Spieler jedoch den Anreiz, strategisch zu kommunizieren. Es gibt zwei potentielle Abweichungen von den Standardannahmen menschlichen Verhaltens, die ehrliche Kommunikation etablieren würden: soziale Präferenzen und kognitive Einschränkungen. Für eine gute Ausgestaltung von Prozessen und Regeln in Komitees ist es entscheidend, die Wichtigkeit dieser Faktoren zu identifizieren. Wenn ehrliches Verhalten auf sozialen Präferenzen beruht und sich somit die Mitglieder auf eine Entscheidung einigen können, gibt es kein Problem für das institutionelle Design. Im Gegensatz dazu könnten kognitive Beschränkungen dazu führen, dass Individuen anstatt ihr optimales Verhalten zu ermitteln naïve Heuristiken benutzen. Dies stellt ein wesentliches Problem dar, da diese Individuen leicht ausgenutzt werden können. Strategisch lügende Individuen könnten die Entscheidungen eines Komitees zu ihren Gunsten beeinflussen, da naïve Individuen die von ihnen verbreiteten unwahrheitsgemäßen Informationen nicht anzweifeln würden.

Um überprüfbare Vorhersagen abzuleiten, verwende ich ein standardökonomisches Modell und zwei verhaltensökonomische Alternativmodelle. Ein Modell beinhaltet soziale Präferenzen, während das andere von einem naïven Wäh-

ler ausgeht, der eine Heuristik verwendet. In einem Experiment teste ich getrennt und gemeinsam Kommunikations- und Abstimmungsentscheidungen, und untersuche, inwiefern sie von der Präsenz heterogener Präferenzen im Komitee abhängen. Die Daten bestätigen keines der obigen Modelle, können aber durch ein Modell kognitiver Heterogenität erklärt werden. Dieses Modell nimmt zwei Spielertypen an. Naïve Level-0-Spieler verhalten sich wie im naïven Wählermodell und intelligente Level-1-Spieler treffen die optimale Entscheidung gegeben der Annahme, dass alle anderen Spieler Level-0 sind. Etwa 80% aller Teilnehmer im Experiment verhalten sich wie Level-0-Spieler; sie sagen die Wahrheit und benutzen eine Entscheidungsheuristik bei der Abstimmung. Die verbleibenden intelligenten Level-1-Spieler lügen strategisch und wenden ihre optimale Entscheidungsregel an. Diese Ergebnisse haben zwei wichtige Implikationen. Erstens warnen die experimentellen Befunde davor, ein niedriges Lügenniveau als ein homogenes Phänomen zu interpretieren, nach dem jeder Teilnehmer generell ehrlich ist und nur selten lügt. Denn, wie die Ergebnisse zeigen, kann solch ein niedriges Niveau auch das Resultat einer kleinen Gruppe von konsistent strategisch lügenden Individuen sein. Zweitens zeigen die Ergebnisse die Notwendigkeit auf, kognitive Heterogenität in theoretische Modelle über Komitee-Entscheidungen zu integrieren und Mechanismen zu finden, die robust gegenüber der Ausnutzung von naïven Spielern sind.

Kapitel 3 wendet den Fokus vom politischen Umfeld des Komitees zu einer grundsätzlichen Frage der Kommunikation: Hängt in strategischen Situationen die Entscheidung für ehrliche Kommunikation davon ab, ob die relevante Information von der Person, die sie bereitstellt (der Sender), selbst erworben oder diese exogen vorgegeben wurde? Zur Beantwortung dieser Frage untersuche ich Kommunikation isoliert von anderen institutionellen Rahmenbedingungen unter Verwendung eines Zwei-Personen-Spiels mit einseitiger Kommunikation. In solch einer Beratungssituation kann der Wunsch, dem Empfänger der Information das eigene Fachwissen zu zeigen, Individuen motivieren, die Wahrheit zu sagen. Dies ergibt sich insbesondere dann, wenn endogene Information kommuniziert wird, d. h. wenn die Genauigkeit der Information von der Fähigkeit des Senders abhängt, den Wahrheitsgehalt aus der gegebenen Information zu entnehmen.

Während in Kommunikationsspielen generell angenommen wird, dass dem Sender die private Information exogen gegeben wird, ist sie in Wirklichkeit oftmals ein Produkt seiner Expertise. Nehmen wir als Beispiel einen Lobbyisten, der die Regierung in Fragen der Bankenregulierung berät. Wenn dieser neue Informationen erhält (z.B. neue Kennzahlen von Banken), muss er die Fähig-

keit besitzen, diese richtig zu interpretieren und die relevante Information zu extrahieren (z.B. über die finanziellen Risiken der Banken). Folglich beinhaltet seine Nachricht nicht nur Informationen über die relevante Frage, sondern auch über seine Kompetenz, und kann daher auch sein soziales Image beeinflussen. Eine Schlussfolgerung ist, dass das Gewicht, welches in der Nutzenfunktionen dem sozialen Image beigemessen wird, von der Art der Expertise, die in der Kommunikation übermittelt wird, abhängt.

Das Kapitel beinhaltet eine experimentelle Untersuchung, die den Effekt von endogener Information auf ehrliche Kommunikation untersucht. In einem Cheap-Talk-Spiel, in dem Individuen komplett gegensätzliche Präferenzen haben, wird der soziale Status der Information variiert. In zwei Experimentalbedingungen erhalten Sender Multiple-Choice-Fragen zu (1) Allgemeinwissen (hoher sozialer Status) und (2) Boulevard-Themen (niedriger sozialer Status). Während die Standardtheorie vorhersagt, dass in beiden Bedingungen der Sender lügen sollte, finde ich, dass in der Bedingung mit hohem sozialen Status signifikant mehr die Wahrheit gesagt wird als in der Bedingung mit niedrigem sozialem Status. Darüber hinaus zeigen die Ergebnisse, dass der wesentliche Kanal für den Effekt die Möglichkeit ist, die eigene Expertise zu signalisieren. Wenn diese Möglichkeit ausgeschlossen wird und der Sender die Information exogen erhält, verschwindet der Unterschied zwischen beiden Bedingungen. Die Ergebnisse zeigen, dass endogene Information den begrenzten Raum für wahrheitsgemäße Informationsübertragung vergrößern kann, der generell in der Standardtheorie für Cheap-Talk-Spiele gefunden wird (siehe für ungenaue Kommunikation, z.B. Austen-Smith, 1993; Battaglini, 2002; Coughlan, 2000).

Während in Kapitel 2 und 3 das Potenzial der Kommunikation zur Informationsaggregation im Fokus steht, widmet sich das Kapitel 4 der Legitimität von politischen Institutionen und deren Effekte auf Regelkonformität und Sanktionierungsverhalten. Die Schaffung von Legitimität ist eine weitere zentrale Funktion von Abstimmungen. Diese ist insbesondere von Bedeutung, wenn die Wähler Entscheidungsgewalt an Autoritäten übertragen, die in ihrem Auftrag regieren und langfristig-orientierte Entscheidungen treffen. Die prozedurale Fairness von politischen Autoritäten kann die institutionelle Legitimität erhöhen, was wiederum die Effektivität der von ihnen umgesetzten Regeln steigern kann (Tyler, 2006).

Der Legitimitätsansatz kann zum Verständnis eines weiteren Phänomens beitragen, das die Standardtheorie nicht erklären kann: Kooperation und Regelkonformität bei Problemen des kollektiven Handelns. Von einer reinen Ab-

schreckungsperspektive aus (Becker, 1968) sollten viele Sanktionen auf Grund von niedriger Nachweiswahrscheinlichkeit eines Vergehens oder niedrigen Strafen rationale Individuen nicht von Trittbrettfahrerverhalten abhalten (z.B. Engel, 2014). Dennoch finden wir in der Realität einen überraschend hoher Grad an Regelkonformität (Nagin, 1998). Erst kürzlich hat die experimentelle Forschung Evidenz für die ursprünglich auf Jean-Jacques Rousseau (1896) und Alexis de Tocqueville (1836) zurückgehende Idee hervorgebracht, dass demokratische Partizipation zu einer höheren Befolgung von Regeln führt (Dal Bó, Foster und Putterman, 2010; Tyran und Feld, 2006).

In Kapitel 4 untersuche ich gemeinsam mit Pedro Robalo und Franziska Tausch die Frage, ob ein unbeteiligter Dritter, der mit Strafautorität ausgestattet ist, in seiner Entscheidung vom institutionellen Verfahren beeinflusst wird. Experimentell wird ein demokratisches Verfahren (Mehrheitswahl) zur Entscheidung über den Einsatz der Strafautorität mit einer Zufallsentscheidung verglichen. Die zugrundeliegende Frage ist, ob die Abstimmung nicht nur das Verhalten derjenigen beeinflusst, die am Verfahren beteiligten sind, sondern auch derjenigen, die verantwortlich für die Umsetzung der Sanktionen sind. In der Realität werden viele Sanktionen von Unbeteiligten verwaltet (z.B. Richter, Polizei). Während die meisten Sanktionen mehr oder weniger etablierten Regeln (Rechtsvorschriften, sozialen Normen, usw.) unterliegen, hat jede Autorität etwas Spielraum bei der Umsetzung der Strafen. Ein bedeutendes Beispiel ist hier die Rechtsprechung: Richter sind an Gesetze gebunden, haben aber in ihren Entscheidungen richterliches Ermessen. Durch diesen Ermessensspielraum bei der Rechtsanwendung besteht die Möglichkeit, dass die Sanktionsentscheidungen unbeteiligter Dritter von der Legitimität des institutionellen Verfahrens beeinflusst werden.

In zwei Experimentalbedingungen variiere ich das institutionelle Verfahren. Vor dem Start einen Gemeinwohlspiels stimmen zuerst alle Teilnehmer ab, ob sie eine Sanktionsautorität (einen unbeteiligten Teilnehmer) einführen wollen. Anschließend wird diese entweder endogen durch ein demokratisches Verfahren mit Mehrheitsregel adoptiert oder exogen durch den Experimentator implementiert. Die Daten zeigen, dass die Bestrafung Trittbrettfahrer effektiver diszipliniert, wenn die Sanktionsautorität endogen aus dem Abstimmungsprozess hervorging. Tatsächlich erwarten die endogen eingesetzten Dritten, dass ihre Bestrafung effektiver ist und wählen daher mildere Sanktionen als ihre exogen eingesetzten Pendants. Das experimentelle Design löst die Probleme der Selbstselektion und der Signalwirkung, die bei Abstimmungen über institutionellen Rahmenbedingungen entstehen können. Im Gegensatz zu Felduntersuchungen

erlaubt dies, die kausale Wirkung der endogenen institutionellen Wahl unabhängig von den Eigenschaften der Wähler und der Sanktionsautorität zu analysieren.

In der Zusammenfassung im Kapitel 5 werden die wichtigsten Ergebnisse der einzelnen Kapitel dargelegt und Implikationen für weitere Forschung und praxisrelevante Anwendungen kritisch diskutiert.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Almost all political institutions seek to resolve some problem of collective choice or collective action. The underlying problem is often caused by a lack of information about other individuals' preferences or private knowledge. In a social dilemma, many individuals would cooperate if they only knew others were willing as well (Fischbacher and Gaechter, 2010). In committees with common values, as envisaged by Condorcet (1785), members would agree on the right decision if the state of the world were known. But when members hold private, and potentially conflicting, information, they may disagree on which decision to take.

This dissertation analyzes the role of communication and voting in group environments where conflicting interests are present, and studies the resulting consequences for institutional design. The focus lies on two topics. Chapters 2 and 3 study the information-aggregation potential of communication, and, in particular, whether non-standard preferences induce truth-telling. Chapter 4 studies the effects of a democratic decision-procedure on sanctioning and compliance behavior.

This dissertation contributes to the recent discipline Behavioral Political Economy which studies the actual behavior of people within political institutions as opposed to traditional approaches relying on assumptions of rational and self-interested actors (Fischbacher, Hillman, and Ursprung, 2015; Schnellenbach and Schubert, 2015). It follows the approach of behavioral and experimental economics by seeking to improve the realism of models through behavioral model extensions and empirical testing (Camerer, 2003). In the

subsequent three independent chapters, I use theory-guided laboratory experiments and integrate behavioral motives into standard models (cognitive hierarchy and social preferences in Chapter 2, image concerns in Chapter 3, legitimacy in Chapter 4). While the standard models of political economy assuming self-interested preferences and full rationality (Rowley, 1993) have yielded substantial insights into political processes, they cannot explain other phenomena, such as overcommunication and compliance. I provide explanations for both phenomena and test these experimentally.

Condorcet (1785) originally introduced the view that political institutions are able to aggregate private information. Its general idea is that voting is used as a communication device: voters reveal their private information through their vote, so-called sincere voting. In particular, the Condorcet Jury Theorem states under which conditions majority-rule voting yields efficient information aggregation through sincere voting; its later extensions reveal the limits of the basic result (Austen-Smith and Banks, 1996; Feddersen and Pesendorfer, 1998; Piketty, 1999). The presence of conflicting interests in groups, for instance, is a factor that typically leads to the failure of sincere voting (Coughlan 2000). In these situations, other non-voting institutions, such as debates, can complement the voting institution to reestablish information sharing.

Whilst communication allows for a more direct form of information aggregation than voting, it can also fail (Coughlan, 2000). This happens in particular in situations without common value, i.e., when actors disagree about their preferred policy. One prominent example is lobbying where informed and interested lobbyists send messages to uninformed, but empowered decision-makers (Crawford and Sobel, 1982). In spite of predicted strategic communication, a large number of experiments find that individuals lie very little, a phenomenon called overcommunication (see for a review Abeler, Nosenzo, and Raymond, 2016).

In Chapters 2 and 3 different forms of overcommunication are studied. In the last decade, the literature has put forward numerous behavioral explanations. These are either based on a particular preference (e.g., preference for being honest, preference for being perceived as honest) (e.g., Ellingsen and Östling, 2010; Fischbacher and Föllmi-Heusi, 2013; Kartik, Ottaviani, and Squintani, 2007; Mazar, Amir, and Ariely, 2008; Sánchez-Pagés and Vorsatz, 2007), a social norm (e.g., individuals like to conform to a social norm of telling the truth)(Diekmann, Przepiorka, and Rauhut, 2015; Rauhut, 2013), or bounded rationality (Cai and Wang, 2006). Some motives may matter more

than others depending on context. In collective choice environments that aggregate many individual decisions, social preferences are likely to play a role. In Chapter 2, I study, together with Mark Le Quement, whether committee members' objective function is not to maximize their own payoffs but rather the sum of individual payoffs, which would induce truth-telling. Alternatively, I also test whether, given the complexity of the decision-making environment, truth-telling is used as a heuristic.

In particular, Chapter 2 studies situations in which committee members face a trade-off between the social goal of making a correct decision and pursuing one's self-interest. In Condorcet Juries where agents hold private information, sharing information truthfully in a deliberation stage can improve the accuracy of the group's decision. Yet in groups with heterogeneous preference types, agents have an incentive to communicate strategically. There are, however, two potential deviations from standard assumptions that may lead to truth-telling: social preferences and cognitive constraints. Identifying these drivers is crucial for designing committees. If truth-telling were the result of individuals exhibiting social preferences and therefore agreeing on the collective decision to take, there would be no problem for institutional design. In contrast, cognitive constraints may lead individuals to use a heuristic instead of calculating optimal behavior, which poses a more substantial problem. Individuals who rely on a truth-telling heuristic could be easily exploited. Sophisticated, strategically lying individuals, could then influence committee decisions in their favor because their lies would be taken at face value by naïve individuals.

To derive testable predictions, I use a standard model and two behavioral alternatives, a social preferences model and a naïve voter model where individuals use a heuristic. Experimentally, I separately and jointly test communication and voting choices and how they depend on the presence of heterogeneous preferences. The data does not confirm any of the above models, but can be explained by cognitive heterogeneity. In particular, it is in line with a cognitive heterogeneity model that assumes two types: naïve level-0 agents behave as in the naïve voter model and sophisticated level-1 agents best respond to the assumption that all others are level-0 agents. I find that roughly 80% of subjects behave as naïve agents, i.e., they tell the truth and use a decision heuristic. The remaining sophisticated agents lie strategically and approximately apply their optimal decision rule. These results have two important implications. First, the findings caution against interpreting low lying rates as an homogenous phenomenon according to which individuals generally truth-

tell and rarely lie, since the lying rates can be equally caused by a small group of sophisticated individuals that consistently lie. Second, they highlight the need to integrate cognitive heterogeneity into theoretical models of committee decision-making and build mechanisms that are robust against exploitation.

Chapter 2 was co-authored with Mark Le Quement from the University of East Anglia. The initial idea was equally developed by both of us. I was leading in developing the experimental design, programming and running the experiment, and analyzing the data. Mark Le Quement was the leading author in the theory sections. The writing was equally shared.

In Chapter 3, I shift the focus from the political environment of a committee to a very basic question of communication: In strategic situations, does truth-telling depend on whether the relevant information was acquired by the sender or exogenously given to her? For this purpose, the chapter isolates communication from other institutional factors within a simple two-person game of unilateral communication. In these situations of advice-giving, the desire to show one's expertise may motivate individuals to truth-tell. This opportunity arises when senders communicate endogenous information, i.e., when the precision of the information depends on the sender's characteristics, more precisely on her ability to extract the true state out of the given information.

Whilst in communication games it is commonly assumed that information is exogenously given to the sender, in reality, the sender's information is often a product of her expertise. Imagine, for instance, a lobbyist who advises the government on regulation policies in the banking industry. When the lobbyist receives new information (e.g., key financial information from banks), she has to possess the necessary skills to adequately interpret the data and extract the true content (e.g., financial risks). Consequently, a message may not only reveal information about a certain state of the world, but also about the sender's expertise and may thus affect her social image. A natural implication is that the weight of social image utility depends on the specific type of expertise transmitted by the information.

Experimentally, this chapter examines the effect of endogenous information on truth-telling in a cheap talk game with completely misaligned preferences and varies the social status of information. In two treatments, senders are provided with multiple-choice questions on (1) general knowledge (high social status) and (2) tabloid topics (low social status). While standard theory predicts the sender to babble in both treatments, I find that truth-telling rates are significantly higher in the high social status treatment. Moreover,

the results clearly show the driving channel to be the ability to signal expertise. When the opportunity to signal expertise is removed and information is provided exogenously to the sender, the difference between high and low vanishes. Thus, the results demonstrate that endogenous information can enlarge the limited scope of truthful information transmission that is typically found in standard cheap-talk games (see for cases of inaccurate communication e.g., Austen-Smith, 1993; Battaglini, 2002; Coughlan, 2000) when senders care about being positively perceived by others. Chapter 3 is single-authored.

While Chapter 2 and 3 have studied the potential of communication to aggregate information, Chapter 4 shifts the focus to the legitimacy of political institutions and its behavioral effects. The creation of legitimacy is another key function of voting procedures. It matters in particular when the electorate defers decisions to authorities that rule on their behalf and administer decisions over time. Procedural fairness of political authorities can increase their institutional legitimacy, which in turn can enhance their success in implementing effective rules (Tyler, 2006).

The legitimacy approach to institutions can contribute to the understanding of another phenomenon that cannot be explained by standard theory: cooperation and compliance in collective action problems. Whereas from a pure deterrence perspective (Becker, 1968), many real-world sanctions should not deter rational individuals from free-riding (whenever sanctions and/or the detection probability are low) (e.g., Engel, 2014), one does observe a surprisingly high degree of compliance (Nagin, 1998). Recently, experimental research has contributed evidence to a long-standing idea, which dates back to Rousseau (1896) and Alexis de Tocqueville (1836), that democratic participation leads to higher levels of compliance with rules (Dal Bó, Foster, and Putterman, 2010; Tyran and Feld, 2006).

In particular, in Chapter 4 I study, together with Pedro Robalo and Franziska Tausch, the question whether third-party punishers are influenced by the institutional choice procedure that implements the sanctioning institution. It compares a democratic institution (majority vote) to a random implementation. The question I tackle is whether voting over sanctions does not only affect the behavior of the parties who take part in the procedure, but also the decisions of the individuals who are responsible for administering them. This question is important to study as most sanctions are administered by third parties (e.g., judges, police). Whilst they are mostly carried out under more or less established sets of rules (social norms, legal rules, etc.), leeway is

granted to the authority by whom it is administered. One prominent example is the judicial system: judges are bound by law but decide cases with some discretion. Therefore, their sanctioning decisions may be affected by the legitimacy of the institutional choice procedure that implemented the sanctioning institution in the first place.

In two treatment I vary the institutional procedure. First, individuals who are players in a public good game vote before the game starts on the introduction of third-party-administered sanctions. Subsequently, the adoption of this institution is either endogenously decided via majority voting reflecting a democratic procedure or exogenously imposed by the experimenter. I find that third-party punishment is considerably more effective at disciplining free riders when punishers emerged endogenously from the voting process. In fact, endogenous punishers anticipate their higher effectiveness and thus choose milder sanctions than their exogenous counterparts. The experimental design addresses the self-selection and signaling effects that arise when subjects can vote on the institutional setting. Thus, our design overcomes the selection problem of field data, i.e., that the causal effect of endogenous institutional choice cannot be disentangled from the individual characteristics of the voters and of the sanctioning authority.

Chapter 4 was co-authored with Franziska Tausch and Pedro Robalo from the Max Planck Institute for Research on Collective Goods. The overall workload was evenly distributed and all co-authors contributed proportionally to all stages of the study.

Chapter 2

Communication and voting in heterogeneous committees: An experimental study

2.1 Introduction

Collective decision-making commonly brings together individuals whose preferences are heterogeneous. Examples include parliamentary committees consisting of members of different political parties or boards of directors consisting of different types of stakeholders (public and private stockholders, employees, etc). Even with heterogeneous preferences collective decision-making often has a common value dimension in the sense that members would agree on the right decision if the state of the world (whether a defendant is guilty or innocent, whether a reform will lower unemployment, whether a job candidate is competent) were known. If the state of the world is instead uncertain, disagreement about the right decision may arise, as different committee members may value the two potential types of error differently (false positive vs. false negative).[1]

[1]For example, there is a large body of empirical evidence showing that jury members hold heterogeneous preferences concerning convicting an innocent vs. acquitting a guilty. Different preferences may be rooted in political attitudes, demographic characteristics, personality traits, etc. In the psychological literature on judicial decision-making there is a strand of literature that empirically studies the effects of demographic and personal characteristics on jury deliberation (for reviews see Devine et al., 2001; MacCoun, 1989; Pennington and Hastie, 1990; Sommers and Ellsworth, 2003).

When heterogeneous preferences are commonly known, standard game theory predicts that rational and self-interested individuals have incentives to misrepresent their private information in debates (Coughlan, 2000). This is wasteful from a welfare perspective, as pooling of information increases the probability of making a correct decision.

One can envisage two potential deviations from standard assumptions that may lead to truth-telling, namely social preferences and cognitive constraints. If, for instance, individuals' objective function is to maximize joint payoffs instead of individuals payoffs, lying incentives are eliminated. A related experimental study (Goeree and Yariv, 2011) with privately known (and potentially heterogeneous) preference types and free-form communication finds that individuals mostly truth-tell and vote in line with the majority of announced signals. The authors rationalize this finding with the presence of social preferences, i.e., subjects behave *as if* they maximize the sum of members' individual payoffs. Cognitive constraints is the second potential deviation from standard assumptions. Many individuals may be unable to communicate strategically or to identify their payoff-maximizing decision rule and instead revert to a simple decision heuristic which involves truth-telling.

Identifying the drivers of behavior is key to good committee design. If truth-telling is the result of de facto homogeneity driven by social preferences enabling truthful communication, there is no institutional design problem. In contrast, the use of heuristics poses a more substantial problem. If individuals are guided by decision heuristics, they may have difficulties identifying the welfare optimal decision rule (i.e., the level of certainty necessary to favor a decision over an alternative).[2] Furthermore, this kind of naïve behavior could be easily exploited. If, for instance, the population consisted of a large group of naïve, truth-telling individuals and a small group of sophisticated, strategically lying individuals, the latter could influence committee decisions in their favor because their lies would be taken at face value.

We study a three-person deliberative jury game (Coughlan, 2000; Feddersen and Pesendorfer, 1998) with two publicly known preference types and majority rule and investigate whether social preferences or cognitive constraints drive the (non-) existence of strategic communication. It is useful to separately analyze the different stages of decision-making. We vary (a) the information provision protocol (public signals vs. private signals) and (b) the committee composition (homogenous vs. heterogeneous). These treatments allow us to

[2]See for a discussion of standards of proof in law and decision theory Schweizer (2016).

observe voting with and without prior communication keeping constant the number of signals that are available in the committee. This comparison is novel in the literature, which has focused on comparing private voting and voting after communication of private signals (Goeree and Yariv, 2011; Guarnaschelli, McKelvey, and Palfrey, 2000). In our experimental design, incentives for strategic communication should be relatively easy to identify. First, the preference misalignment is particularly large and salient. Subjects have fixed types and know each others' preference type throughout the game. Second, communication takes place in the form of straw votes (pre-defined messages) that minimize the scope for the emergence of social preferences through communicative interaction.

To derive testable predictions, we use the standard model of own payoff-maximizing strategic agents (Coughlan, 2000; Feddersen and Pesendorfer, 1998; Le Quement and Yokeeswaran, 2015) and two behavioral alternatives. The first alternative is a social preference model of joint payoff maximization, as mentioned in Goeree and Yariv (2011). The second alternative is a naïve voter model where all individuals truth-tell and use a decision heuristic, i.e., vote in line with the majority of announced signals (*majority heuristic*). This model is in the spirit of Condorcet (1785) who assumed jury members to act non-strategically and prefer the alternative that is most likely to be correct.

On aggregate, our results clearly reject the standard model and both behavioral alternatives: (a) We find no evidence of babbling as predicted by the standard model; (b) the preference profile of the committee (heterogeneous vs. homogenous) does not affect individuals' voting behavior as predicted by the social preference model; (c) the two different preference types do not use the same decision rule as predicted by the naïve voter model.

Disaggregating results, we find a large heterogeneity across individuals. In order to identify different cognitive types who are potentially present, we introduce a simple model of level-k thinking (see for instance Crawford and Iriberri, 2007; Nagel, 1995; Stahl and Wilson, 1994, 1995). Level-0 agents behave as in the naïve voter model. Level-1 agents best respond to the assumption that all others are level-0 agents. The former lie when they hold a signal that is contrary to their preference bias and apply their optimal decision rule at the voting stage. Our findings indeed suggest that subjects can be categorized into these two groups. In particular, the vast majority of subjects (82%) consistently truth-tells. These subjects use their type-specific optimal decision rule in only 24% of all cases. In contrast, 18% of subjects consistently lies after a

contrary signal. The latter subjects furthermore use their type-specific optimal decision-rule in 60% of the time. Consistent lying is thus strongly associated with applying the type-specific optimal decision rule.

In the real world, committees are usually much larger than the three-person committees studied in our experiment. Even if sophisticated agents are relatively rare, the law of large number implies that the larger a committee, the higher the likelihood that it will contain at least some sophisticated agents. Our results suggest that sophisticated agents have a strong impact on outcomes because their lies are taken at face value. In fact, we find the payoff gain from lying after a contrary signal to be about 26%.

We proceed as follows. Section 2.2 presents a brief overview of the literature. Section 2.3 presents our experimental design and Section 2.4 the theoretical predictions. Section 2.5 analyzes aggregate behavior with an eye to testing the respective predictions. Section 2.6 focuses on heterogeneity in behavior, presents the cognitive heterogeneity model and analyzes its predictions. Section 2.7 concludes.

2.2 Related Literature

Building on Condorcet's seminal essays on voting (see Condorcet, 1785), a theoretical literature that models voting as information aggregation has blossomed over the last two decades. Early contributions (Austen-Smith and Banks, 1996; Feddersen and Pesendorfer, 1998) study private voting (see also Feddersen and Pesendorfer, 1996; Gerardi, 2000; Martinelli, 2006; Meirowitz, 2007; Persico, 2004). Key findings have been confirmed and qualified experimentally in Guarnaschelli, McKelvey, and Palfrey (2000), Esponda and Vespa (2014), Grosser and Seebauer (2016), Battaglini, Morton, and Palfrey (2008), and Battaglini, Morton, and Palfrey (2010).

A set of newer contributions study the case of voting preceded by communication and have focused on the truthful-sincere equilibrium. A milestone is the negative result obtained by Coughlan (2000) for the case of publicly known heterogeneous preference types: If full truth-telling leads to disagreement, then there exists no truthful-sincere equilibrium. Le Quement and Yokeeswaran (2015) provide an equilibrium prediction for such committees under unanimity rule. Deimen, Ketelaar, and Le Quement (2015) offer a complementary analysis that assumes conditionally correlated signals. A parallel research

agenda has studied the extent to which uncertainty about preference types affects the possibility of communication (Austen-Smith and Feddersen, 2006; Le Quement, 2013; Meirowitz, 2007; Van Weelden, 2008).

Voting with communication has also been examined experimentally.[3] Guarnaschelli, McKelvey, and Palfrey (2000) study a homogeneous jury and find, in contradiction with the intuitive prediction of full truth-telling, a 5% lying rate and skepticism towards information provided by others. Goeree and Yariv (2011) study the case of privately known (and potentially heterogeneous) preference types with free-form communication. The authors' confirm the theoretical prediction formulated in Gerardi and Yariv (2007), namely that all voting rules are equivalent given unrestricted communication. Furthermore, they find that subjects on average follow a simple heuristic which consists of truth-telling and subsequently voting with the majority of announced signals.

2.3 Experimental Design and Theoretical Predictions

We here describe the treatments and the experimental procedure.

2.3.1 Treatments

Our main treatment is a Condorcet Jury Game with private information and heterogeneous preferences. A committee, composed of three subjects, has to choose between two alternatives by majority vote. If the state of the world were known, each preference type would favor choosing the alternative that matches the true state of the world. However, different preference types value the two potential errors differently. The state of the world is not observable, but takes one of two possible values, both being ex ante equally probable. Specifically, the state is the (red or blue) jar selected by nature and the decision is either

[3] Our focus, as well as that of the here reviewed literature, is on deliberation as information aggregation. We refer to Hafer and Landa (2007) and Dickson, Hafer, and Landa (2008) for theoretical and experimental work on deliberation modeled as a rational and strategic process of self-discovery.

red or *blue*.[4]

The timing of the game is as follows. There is an information stage (stage 1) at which information regarding the color of the jar is received and exchanged. At stage 2, each subject casts a vote from the set $\{red, blue\}$ and a collective decision is made. In stage 3, subjects observe the number of votes for each jar, the committee decision, the jar selected by nature as well as their payoffs.

We vary the main treatment on two dimensions: (a) the preferences of subjects and (b) the information protocol, see Table 2.1 below.

Table 2.1: Overview of treatments

		Preferences	
		Homogeneous	Heterogeneous
Information	Private signals	Private-Hom	Private-Het
	Public signals	Public-Hom	Public-Het

Preferences (Heterogeneous vs. Homogenous): In heterogeneous (*Het*) committees there are two possible preference types, *red* or *blue*, whose payoffs depend on the group decision and the realized jar (see Table 2.2). As can be seen from the table, red (blue) types are biased towards the red (blue) jar. If agents are risk neutral and self-interested, this payoff specification is equivalent to the preference specification introduced in Feddersen and Pesendorfer (1998) and Coughlan (2000).[5] Committee composition is common knowledge at the start of the game. In each committee there are either two *red* types and one *blue* type or vice-versa.[6] In contrast, homogeneous (*Hom*) committees consist only of one preference type, either all subjects are *red* or *blue*.

[4]We follow Guarnaschelli, McKelvey, and Palfrey (2000) and Goeree and Yariv (2011) in adopting a neutral description. In the jury interpretation, the group chooses between convicting and acquitting a defendant who is either guilty or innocent.

[5]In those models, a juror's payoff is determined by a commonly known parameter $q \in (0, 1)$. He obtains payoff $-q$ (resp. $-(1-q)$) if the chosen jar is red (blue) while the realized jar blue (red). Payoffs from choosing the correct jar are normalized to 0. We exclude negative payoffs by applying a positive transformation to the original ones. Payoffs in Table 2.2 are equivalent to $q_{blue} = \frac{5}{6}$ for blue types and $q_{red} = \frac{1}{6}$ for red types in the original specification.

[6]A subject whose preference type is (not) shared by some (any) other subject is called a majority (minority) subject.

Table 2.2: Payoff structure

Group Decision		True Jar: Blue Jar	True Jar: Red Jar
	Red	10	40
	Blue	160	10

Blue Type

Group Decision		True Jar: Blue Jar	True Jar: Red Jar
	Red	10	160
	Blue	40	10

Red Type

Information (Private vs. Public): In *private* treatments, information is transmitted in two stages. In substage 1.a, each agent privately observes a signal. A signal takes the form of a red or blue ball randomly drawn with replacement from the realized jar. The blue (red) jar contains 7 (3) blue balls and 3 (7) red balls.[7] In the subsequent substage 1.b, each agent picks a simultaneously observed public message from the set $\{red, blue\}$.[8] In other words, the *private* treatments feature a round of simultaneous cheap talk communication. Messages are shown with an indication of the subject's preference type, but without a player identifier. In so-called *public* treatments, information comes in the form of three i.i.d. public signals, which is equivalent to forced sincere communication.

We introduce the following two simplifying notations. First, we shall repeatedly be referring to the *observed signal profile* of a subject in stage 1. In *private* treatments, it corresponds to a subject's own signal combined with the two signals announced by others. In *public* treatments, it corresponds to the three public signals observed. Second, given that types are symmetric, we call a red signal held by a red type a *conform* signal and a blue signal held by a red type a *contrary* signal. Equivalently, we call decision *red* (*blue*) the *conform (contrary)* decision for a red type (and vice versa for blue types). As can be seen in Table 2.2, payoffs are symmetric across red and blue types. Thus, we should expect identical behavior by red and blue types at symmetric information sets. To see that, consider an outcome given by a profile of signals combined with a decision. Construct the symmetric outcome, which is obtained by replacing any blue (red) signal by a red (blue) signal as well as reversing the decision. The expected payoff of a red (blue) type given the first

[7]Formally, a signal s is an independent Bernoulli trial from a state-dependent distribution with $P(s = red \mid red) = P(s = blue \mid blue) = p = 0.7$, while $P(s = blue \mid red) = P(s = red \mid blue) = 1 - p = 0.3$.

[8]Note that a subject does not have the possibility to refrain from sending a message, as in the original Coughlan (2000) setup.

outcome is the same as that of the blue (red) type given the second outcome.

2.3.2 Experimental Procedure

We use a between-subjects design. Each session contains the following parts: (1) treatment, (2) strategic communication test (SCT) (only *private*), (3) individual decision test (IDT), (4) lying aversion test and (5) social value orientation test. Payoffs from each of the post-experimental tests are learned after the last test. While (4) and (5) are standard tests adopted from the literature, (2) and (3) are introduced by us. See Appendix 2.8.1 for a description of the post-experimental tests (2)-(5).

At the start of each treatment, subjects are randomly assigned a preference type and a matching group of 6 subjects.[9] In each period, two three-subject committees are randomly formed. An equal number of subjects is assigned to each preference type. In *Hom* treatments, each matching group contains either only blue or only red types. In *Het* treatments, each matching group contains three blue and three red types. The game is played repeatedly over 20 rounds with random rematching within each matching group. In *Het* treatments, each subject is thus very likely to experience multiple rounds in minority and in majority.

The experiment was conducted in the BonnEconLab in February and March 2015. It was programmed and conducted with the software z-Tree (Fischbacher, 2007) and organized with the software hroot (Bock, Baetge, and Nicklisch, 2014). A total of 384 University of Bonn students from various disciplines (15% with an economics major) participated in 16 sessions (each of 24 subjects). 96 subjects participated in each treatment, yielding 16 independent matching groups per treatment. Subjects received written instructions which were read out loud by the experimenter (see Appendix 2.8.3 for an English transcript of the original German instructions). To familiarize subjects with the game and ascertain that they understood it fully, we asked control questions that had to be answered correctly. Subjects were given the opportunity to privately ask questions. The amounts earned from the experiment were exchanged at a rate of 150 ECU = 1 Euro. Subjects received the payment from all 20 rounds, which averaged 10.50 Euros and ranged from 5.50 Euros to 16.50 Euros. Subjects additionally earned an average of 4.68 Euros in the

[9]Subjects are not informed about the size of the matching group.

post-experimental tests. On average, one session lasted 65 minutes (40 minutes jury experiment and 25 min post-tests). 58.6 % of subjects were female and average age was 22.6 years.

2.4 Theoretical Predictions

This section introduces the standard model, the social preference model of joint-payoff maximization and the naïve voter model. The first two models assume rational and risk-neutral agents while they differ on the assumed preferences, i.e., agents maximize own payoffs in the standard model and joint payoffs in the social preference model. We focus on equilibria in symmetric strategies, in which agents with identical payoff functions use the same strategy. For each treatment, we derive theoretical predictions from each of the three models. Subsequently, we state a set of testable hypotheses concerning differences in outcomes across treatments.

2.4.1 Standard Model

The standard model is analyzed in Feddersen and Pesendorfer (1998) and Coughlan (2000). It assumes that agents only maximize own expected payoffs and are risk-neutral. Given the payoffs in Table 2.2, an agent favors the conform decision if the conform jar has a conditional probability of at least $\frac{1}{6} \approx 0.167$. The conditional probability of a conform jar after 0, 1, 2 and 3 conform signals is $.07, .3, .7$ and $.93$, respectively. The optimal decision rule of each preference type is thus to choose the conform decision if at least one of the three signals is conform. We denote by $\Lambda(x)$ the decision rule specifying the probabilities of picking the conform decision after r conform signals:

$$[\Lambda(x)](r) = \begin{cases} 0 & \text{if } r = 0, \\ x & \text{if } r = 1, \\ 1 & \text{if } r \geq 2. \end{cases}$$

The rule $\Lambda(1)$ is thus the optimal decision rule of each type. According to the impossibility result obtained by Coughlan (2000) for a game featuring a vote preceded by simultaneous cheap talk, there exists no equilibrium in which all agents truth-tell and vote sincerely if the committee contains at least one

blue-biased and one red-biased agent. To understand the result, assume that the committee contains a simple majority of blue-biased agents. The decision rule applied in the above putative equilibrium is the optimal decision rule of blue-biased agents, i.e., choose red only if three red signals are observed. At the communication stage, the red-biased agent acts under the assumption that his announcement is pivotal (i.e., affects the final outcome) and thus infers that the two other agents hold a red signal. This in turn implies that he favors a red decision. If he holds a blue signal, he thus deviates to announcing a red signal. Our equilibrium prediction for each of the treatments is given below. For *private* treatments, we focus on equilibria featuring maximal information sharing.

Proposition 2.1.
a. Private-Het: Majority agents truth-tell while the minority agent babbles. Majority agents condition their vote only on majority agents' signals. They vote for the conform decision unless they jointly hold two contrary signals. The minority agent conditions his vote on all members' signals and applies $\Lambda(1)$ *to the observed signal profile.*
b. Private-Hom: All agents truth-tell. All agents apply $\Lambda(1)$ *to the observed signal profile.*
c. Public-Het: All agents apply $\Lambda(1)$ *to the observed signal profile.*
d. Public-Hom: All agents apply $\Lambda(1)$ *to the observed signal profile.*

The intuition for our prediction for the *Private-Het* treatment is as follows. At the voting stage, the optimal decision rule of the majority preference type conditional on two signals is implemented. This involves choosing the conform decision unless the two signals are contrary. The minority agent is never pivotal at the voting stage and is thus indifferent between both voting decisions. At the communication stage, a majority agent recognizes that his optimal decision rule is implemented given the publicly pooled information. A majority agent's announcement is pivotal at a unique signal constellation which encourages truth-telling. Assume that red-biased agents are the majority and consider a red-biased agent i. The unique pivotal scenario is when he holds a red signal and others hold blue signals. Announcing a red (blue) signal leads to a red (blue) decision. Indeed, while the voting decision of agent i and the minority agent is independent of i's announcement (the first votes red, the other one blue), the other red-biased agent only votes red if i announces red. Clearly, i prefers to truth-tell. On the other hand, the communication incentives of a minority agent are trivial. Given that his announcement is ignored, he is

indifferent between all messages and accordingly has no incentive to deviate from babbling. As to *Private-Hom*, note that truth-telling is trivially incentive compatible as an agent knows that his optimal decision rule is implemented at the decision stage given pooled information.

We derive three treatment hypotheses from the above proposition concerning (a) communication, (b) use of information and (c) voting behavior.

Hypothesis 2.1.
a. Communication: With private information and communication, communication by minority subjects in heterogeneous committees is less informative than (i) by majority subjects in heterogeneous committees and (ii) by subjects in homogenous committees.
b. Use of information: With private information and communication, majority subjects condition their vote less on the announcement of minority subjects than on that of majority subjects.
c. Voting: With public information, the frequency of a conform vote given an observed signal profile containing one conform signal is the same in homogeneous and heterogeneous committees.

The hypothesis concerning the voting rule focuses on *public* treatments because these by definition exclude any potential skepticism towards information arising as a consequence of communication. These treatments thus provide clean evidence of how subjects decide on the basis of unambiguously trustworthy public information. We focus on behavior given a single conform signal because we expect most of the variation in behavior (across subjects or treatments) to happen at this particular information set.

2.4.2 Social Preference Model

In this model we assume that agents maximize the sum of committee members' individual type-specific payoffs (following the observation of Goeree and Yariv (2011)).[10] Agents thus behave as if they all shared the same payoff function

[10]The behavioral literature proposes different explanations for group-induced preferences (e.g., social preferences, altruism, social norms) as well as different approaches to modeling these preferences. The literature on social preferences features outcome-based models that focus on inequity aversion or taste for efficiency, as well as intention-based models that highlight the role of reciprocity, kindness, etc. See for example Bolton and Ockenfels (2000), Charness and Rabin (2002), Fehr and Schmidt (1999), and Rabin (1993).

given by the average payoff function. In a committee with two (one) blue agents and one (two) red agent, this implies that agents require a conditional probability of the red jar of approximately 0.61 (.39) in order to favor the red decision. Accordingly, the optimal decision rule conditional on three signals is to vote in line with the majority of signals ($\Lambda(0)$). We obtain the following equilibrium predictions. For *private* treatments we focus on equilibria featuring maximal information pooling, as in our analysis of the standard model.

Proposition 2.2.
a. Private-Het: All agents truth-tell and apply $\Lambda(0)$ *to the observed signal profile.*
b. Private-Hom: All agents truth-tell and apply $\Lambda(1)$ *to the observed signal profile.*
c. Public-Het: All agents apply $\Lambda(0)$ *to the observed signal profile.*
d. Public-Hom: All agents apply $\Lambda(1)$ *to the observed signal profile.*

The above model thus predicts truth-telling for any committee composition. Committee composition however affects the implemented decision rule. While heterogeneous committees vote in line with the majority of signals, homogenous committees implement the type-specific decision rule $\Lambda(1)$. Note that our model corresponds to the extreme point of a continuum of models in which a parameter (say $\alpha \in [0, 1]$) measures the degree of altruism of agents. Agents maximize a function given by α times their individual payoff and $1 - \alpha$ times the total committee payoff. We set $\alpha = 0$ for simplicity of exposition, but our predictions for all the treatments would still hold for α small enough.[11] We derive the following set of hypotheses from the above proposition.

Hypothesis 2.2.
a. Communication: With private information and communication, communication is equally informative (i) in homogeneous and heterogeneous committees, and (ii) across majority and minority subjects.
b. Use of information: With private information and communication, all announced signals are equally used by all subjects.
c. Voting: Subjects apply different decision rules depending on whether they are in a homogeneous or heterogeneous committee. The frequency of a conform

[11] Namely, $\alpha \leq 0.699$ for minority *Private-Het* subjects and $\alpha \leq 0.402$ for majority *Private-Het* subjects. While the specific utility function assumed allows us to generate point predictions, other forms of social preferences, as inequity aversion Fehr and Schmidt (1999) or a taste for efficiency Charness and Rabin (2002) would also predict treatment differences.

vote given an observed signal profile containing one conform signal is higher in homogenous committees than in heterogeneous committees.

2.4.3 Naïve Voter model

Condorcet (1785) originally assumed that all individuals have the same objective of making a correct decision, ignore the strategic aspects of committee-decision making and simply vote as if they were the only voter. Translating this idea into our setting means that agents truth-tell and vote in line with the majority of announced signals. They thus choose the alternative that is most likely to be true without taking into account their expected payoffs. This naïve voter model is also in line with the behavior of subjects in treatments with communication in the experiment of Goeree and Yariv (2011).

Proposition 2.3.
a. Private-Het: All agents truth-tell and apply $\Lambda(1)$ *to the observed signal profile.*
b. Private-Hom: All agents truth-tell and apply $\Lambda(1)$ *to the observed signal profile.*
c. Public-Het: All agents apply $\Lambda(1)$ *to the observed signal profile.*
d. Public-Hom: All agents apply $\Lambda(1)$ *to the observed signal profile.*

Hypothesis 2.3.
a. Communication: With private information and communication, communication is equally informative (i) in homogeneous and heterogeneous committees, and (ii) across majority and minority subjects.
b. Use of information: With private information and communication, all announced signals are equally used by all subjects.
c. Voting: Subjects apply the same decision rule independent of (i) whether they are in a homogeneous or heterogeneous committee and (ii) whether they have a blue bias or a red bias.

2.5 Results: Aggregate behavior

In what follows, we analyze communication, the use of information and voting behavior and test the hypotheses formulated in our theoretical predictions

section. We pool red and blue types.[12]

2.5.1 Communication

Table 2.3 shows average lying rates based on individual averages conditional on the signal received for *Private-Hom* subjects, minority *Private-Het* subjects and majority *Private-Hom* subjects. The lying rate after a conform signal is approximately 0 for all three types of subjects. On the other hand, the lying rate after a contrary signal is substantially larger for all three types, though it remains low in absolute terms.

Table 2.3: Lying rates in *private* treatments in %

Signal	Private-Hom	Private-Het
contrary	10.2	
contrary in minority		21.9
contrary in majority		14.9
conform	0.7	
conform in minority		1.0
conform in majority		0.5

We find no evidence of babbling by minority *Private-Het* subjects. Given that they have a lying rate of almost 0 after conform signals, babbling would imply that they virtually always lie after a contrary signal. A one-sided t-test clearly rejects this conjecture ($p < 0.001$). For this and all following tests, we use matching groups as independent units of observation. However, we find marginally significant evidence that minority *Private-Het* subjects lie more after contrary signals than majority *Private-Het* subjects (one-sided Wilcoxon signed-rank, WX test, $p = 0.098$) and significant evidence that minority *Private-Het* subjects lie more than *Private-Hom* subjects (one-sided Mann-Whitney, MW test, $p = 0.02$). The increased lying rate of minority

[12] As already noted earlier, this should be unproblematic given the symmetry of payoffs across types. This is confirmed by statistical analysis. For each type of signal held (conform or contrary) and each possible committee position (i.e., *Private-Het* majority, *Private-Het* minority or *Private-Hom*), we do not find significant differences between voting and communication decisions (two-sided Mann-Whitney rank-sum test) (MW test in what follows).

subjects is in line with the idea of the unilaterial deviation scenario arising in the hypothetical truthful-sincere equilibrium analyzed by Coughlan (2000), i.e., minority types lie to majority types hoping that their message is taken at face value, which in turn would bend the majority types' decision rule towards the one of the minority type.

Behavior at the communication stage thus yields mixed results. There is clearly no evidence of babbling by minority subjects as predicted by the standard model. On the other hand, although we find a large degree of truth-telling in both homogenous and heterogeneous committees as predicted by both alternative models, minority subjects in *Private-Het* lie significantly more than subjects in *Private-Hom*, which is at odds with both models.[13]

Result 2.1.
Subjects to a large extent truth-tell. But there is marginally more truth-telling in homogenous committees than in heterogeneous committees.

2.5.2 Use of information

The standard model predicts that majority subjects in heterogenous committees ignore the information provided by minority types. Both alternative models, however, predict that all announced signals are used equally by all subjects (as everybody is predicted to truth-tell). We therefore focus on the relevant case of majority *Private-Het* subjects and analyze the extent to which they condition their voting decision on the announcement of minority *Private-Het* subjects. Table 2.4 shows a majority type's frequency of choosing the conform decision as a function of his own signal (a conform signal takes the value of 1 and a contrary signal takes the value of 0) and the announcement of the two remaining subjects, one majority type and one minority type. Choice frequencies show that the information provided by the minority type is influential. To see that, compare choice frequencies in cases that differ only according to the message announced by the minority type: 1 vs 3, 2 vs 4, 5 vs 7 and 6 vs 8. Choice frequencies furthermore show that minority type announcements are approximately as influential as majority type announcements (compare cases 2 vs 3 and 6 vs 7).

[13]We refer to Appendix 2.8.2 for a short analysis of the impact of lying aversion (as measured in the post-experiment lying aversion test) on lying behavior in the treatments. In sum, lying aversion has little explanatory power overall, but strongly correlates with the lying behavior of the subsample of subjects who lied at least once.

Table 2.4: Voting behavior by majority types in Private-Het

Case	Observed signal profile			Voting behavior		WX p-value
	own	other majority	other minority	predicted	actual	
1	0	0	0	0	0.04	
2	0	1	0	1	0.17	0.196^1
3	0	0	1	0	0.12	
4	0	1	1	1	0.92	
5	1	0	0	1	0.44	
6	1	1	0	1	0.97	0.046^2
7	1	0	1	1	1.0	
8	1	1	1	1	1.0	

Notes: The observed signal profile includes the majority type's *own* signal, the message by the other *majority* type and the message by the *minority* type. The voting behavior shows the *predicted* frequency to vote conform (according to the standard model) and the *actual* frequency to vote conform. WX is a Wilcoxon signed-rank test ([1]one-sided,[2]two-sided). The unit of independent observation is the matching group.

The above findings are confirmed by statistical analysis. We perform two comparisons that each hold the observed signal profile constant. The first comparison involves cases 2 and 3, for which the respective theoretical predictions differ according to the standard model. Both cases involve an observed signal profile containing only one conform signal. When the conform message is sent by the other majority type (Case 2), the prediction for the majority type is to vote conform. When the same message is however sent by the minority type (Case 3), the prediction is a contrary vote. We find that the frequencies of a conform decision do not differ between cases 2 and 3 (one sided-WX test, $p = 0.196$). The second comparison involves cases 6 and 7, for which the standard model predicts the same frequency of conform votes. In cases 6 and 7, the majority type holds a conform signal, which implies that he favors a conviction independently of the signals announced by others. A two-sided WX test rejects ($p = 0.046$) the hypothesis that the frequency of a conform decision in case 6 is equal to that in case 7. Here, our statistical analysis reveals that a minority type's announcement is actually even more influential than that of a majority type. A potential explanation is that a minority type announcing a

signal that contradicts his bias (e.g., a blue-biased subject announcing a red signal) is naturally perceived as credible. Intuitively, the suspicious scenario is rather that of a minority type announcing a signal that confirms with his bias. Information sent by minority subjects is thus not disregarded. This result rejects the prediction of the standard model, but is in line with both behavioral models.

Result 2.2.
In heterogeneous committees majority subjects condition their vote on the announcement of the minority subject. Majority subjects condition their vote on the announcement of the minority subject approximately as much as on that of a majority subject.

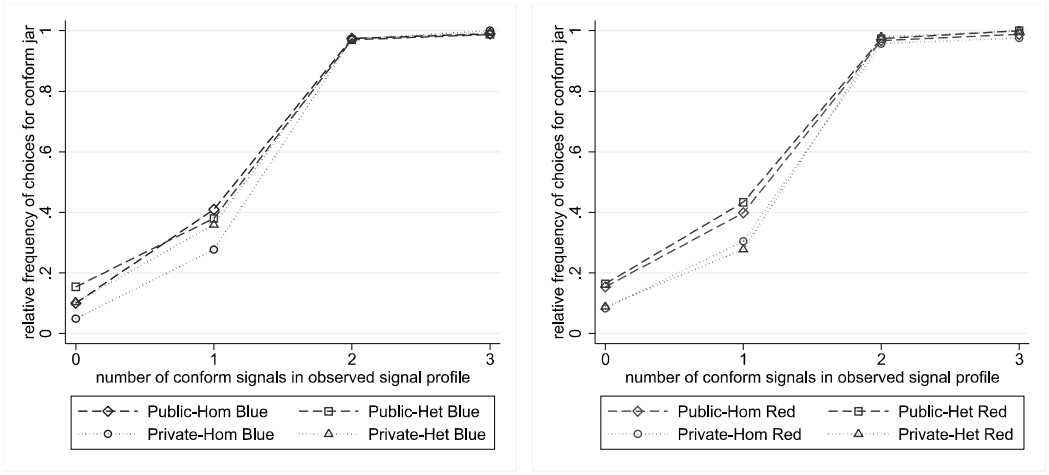
2.5.3 Voting behavior

Figures 2.1a and 2.1b show, for each treatment, the frequencies of votes for the conform jar of each preference type as a function of the number of conform signals in the observed signal profile. For all treatments and preference types, subjects vote conform with a probability that is clearly smaller than one (around 0.35) given a unique conform signal, which is also confirmed by statistical tests. A t-test rejects that the frequency of a conform vote given one conform signal is equal to 1 in the *public* treatments (one-sided t-test, $p < 0.001$). Average decision rules thus exhibit a reversal to the middle.

The social preference model predicts for the heterogeneous committee in the *public* treatment that the frequency of a conform vote given one conform signal is equal to 0, which is also rejected by a one-sided t-test ($p < 0.001$). In other words, subjects do not apply $\Lambda(0)$ to the observed signal profile either. More importantly, the frequency of a conform decision given one conform signal does not differ between homogenous and heterogeneous committees in the *public* treatments (two-sided MW, $p = 0.89$), which means that voting decisions do not depend on the group composition, which the social preference model would have predicted.

The naïve voter model predicts that subjects vote with the majority of announced signals, i.e., vote red if there are at least two red signals (and vice-versa for blue), *independent* of their own preference type. To test this prediction, we look at the voting decision *conditional on red signals* in the observed signal profile (see Figure 2.2). Both after one and after two red

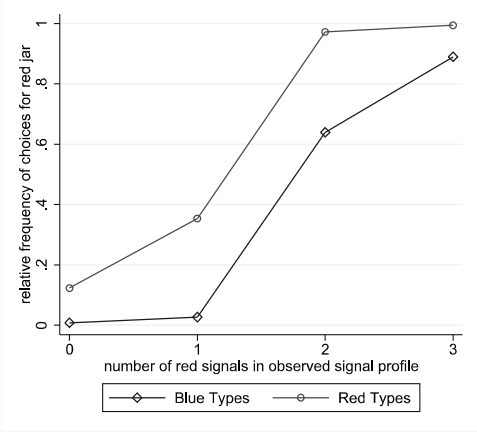
Figure 2.1: Frequencies of choices for the conform jar



(a) Blue types (b) Red types

signals, the probability of a red vote by a red type is much larger than by a blue type (two-sided MW, $p < 0.01$, in each treatment).

Figure 2.2: Frequencies of choices for the red jar across types



Notes: This figure displays voting behavior for red and blue types pooled over treatments.

Result 2.3.
Voting decisions in public treatments do not depend on the group composition. Both preference types apply as decision rule roughly $\Lambda(.4)$*, which implies that*

blue types use a significantly different decision rule than red types (conditional on the number of red signals).

We briefly add a comment on risk aversion. If subjects had very concave utility functions and were thus very risk averse, the utility maximizing decision rule would be $\Lambda(0)$ given three signals. To test whether risk attitudes influenced behavior, we run a regression for the *public* treatments where the dependent variable is a dummy equal 1 if the subject votes for the conform decision and 0 otherwise. The coefficient for risk aversion is marginally significant ($p \leq 0.10$) and small in size, in contrast to the coefficients for the IDT threshold and the dummies for the number of conform messages. We therefore conclude that risk aversion had a negligible impact on behavior (see regression in Appendix 2.8.2).

2.5.4 Summarizing findings

The standard model is clearly rejected. We find no evidence of babbling by minority subjects in *Private-Het*. Second, majority subjects condition their vote as much on minority subjects' announcements as on those of majority subjects. Third, in *public* treatments, subjects' average decision rule is not $\Lambda(1)$. Instead, the average decision rule is heavily skewed towards the majority heuristic. The social preference model is also clearly contradicted. We observe no significant shift in decision rules between heterogeneous and homogenous committees in the public treatments, which indicates that committee composition does not significantly affect subjects' objective function. Finally, the naïve voter model is also not confirmed by the data. Blue and red types do not use the same decision rule. Aggregate lying rates are low, but there is evidence of increased lying rates by minority subjects in heterogeneous committees (albeit at a low level).

2.6 Heterogeneity in behavior

Our previous section on aggregate behavior documents a low lying rate and an intermediate decision rule that lies in between the type-specific optimal decision rule and the majority heuristic. In the following section, we analyze whether the observed aggregate behavior reflects homogenous individual

behavior or rather covers individual heterogeneity. To guide the empirical analysis, we propose a cognitive heterogeneity model with different cognitive types. We subsequently test the predicted type classification.

2.6.1 Cognitive Heterogeneity Model

Heterogenity in behavior has been modeled by level-k reasoning models (see Crawford and Iriberri, 2007; Nagel, 1995; Stahl and Wilson, 1994, 1995) which focus on the interaction between agents whose depth of reasoning, as captured by an integer k, is heterogeneous.[14] In the standard version, a level-k thinker best responds to the assumption that all other agents are level-$(k-1)$ agents. The strategy used by level-0 agents is exogenously specified and the behavior of remaining agents is thus characterized recursively. Experimenters have found that for a variety of games (see Kawagoe and Takizawa (2012) for centipede games, see Crawford and Iriberri (2007) for auctions), given distributions of level-k types fit the data quite well. Level-k has also been used to describe behavior in cheap talk games (Cai and Wang, 2006; Wang, Spezio, and Camerer, 2009) where the level-0 strategy of a sender is assumed to be truthful.

Assume the following simple specification of the level-k model. Level-0 agents are assumed to behave as in the naïve voter model: they truth-tell and vote for the decision indicated by the majority of announced signals. Level-1 agents best respond to the assumption that all others are level-0 agents and maximize individual payoffs. This involves lying after a contrary signal not only in minority but also in majority (whether in a heterogeneous or in a homogenous committee), as well as applying the type-specific payoff-maximizing decision rule. The lying incentive in minority echoes the profitable unilateral deviation scenario arising in the hypothetical truthful-sincere equilibrium analyzed by Coughlan (2000). Lying bends the decision rule applied by majority types (if the lie is taken at face value) and is therefore a profitable individual deviation. In majority and in a homogenous committee lying can bend the decision rule of level-0 types who wrongfully apply the majority heuristic instead of the optimal decision rule. This is only a profitable deviation if there is a high fraction of level-0 types. Interestingly, lying in majority is benevolent as this improves the expected payoff of all subjects sharing the same preference type.

[14] See also Goeree and Holt (2004) for a related model of noisy introspection.

The above model ignores important behavioral features that are presumably empirically relevant. First, lying in majority is less intuitive than lying in minority. Second, agents act noisily in responding to beliefs, as captured for example by the popular Quantal-Response model proposed in McKelvey and Palfrey (1995, 1998). We propose a noisy version of the above introduced model of level-k thinking with level-0 and level-1 agents that differ in their strategic sophistication. This generates two main predictions: First, subjects who lie both in minority and majority have a higher sophistication level that those who only lie in minority. Second, a higher sophistication level is associated with a lower propensity to make errors.

Let any level-1 agent exhibit a sophistication level s drawn from a distribution g with full support on $[0, 1]$. Variable s determines the propensity of a level-1 agent to make errors. More precisely, let any s be associated with probabilities $l(z, s)$ and $d(s)$. The function $l(z, s)$ indicates the probability that a level-1 agent of sophistication level s lies after a contrary signal given that a total of $z \in \{1, 2, 3\}$ agents (him included) have his preference type in the committee. The function $d(s)$ indicates the probability that the level-1 agent applies decision rule $\Lambda(1)$ to the observed signal profile as opposed to $\Lambda(0)$, at the voting stage. We make the following extra assumptions. First, the four above introduced functions are continuous and monotonically increasing in s, reflecting the fact that more sophisticated agents are less prone to make mistakes. Second, $l(z, 1) > .5, \forall z \in \{1, 2, 3\}$ and $d(1) > .5$, capturing the fact that a maximally sophisticated level-1 agent is more likely than not to act optimally, whatever the committee composition. Third, $l(1, s) > l(2, s) > l(3, s), \forall s \in [0, 1]$, reflecting the fact that lying is more intuitive the fewer agents share one's preference type. To close the model, we assume that level-1 agents always truth-tell after a conform signal. We assume that the committee only contains level-0 and -1 agents and therefore do not describe the behavior of higher order types. We summarize our prediction for the above introduced noisy level-k thinking model in the following proposition.

Proposition 2.4.
a. Private-Het and Private-Hom: Level-0 agents truth-tell and vote for the decision indicated by the majority of signals in the observed signal profile. Level-1 agents truth-tell after a conform signal. A level-1 agent of sophistication s applies $\Lambda\left(d(s)\right)$ *to the observed signal profile.*
b. Private-Het: After a contrary signal, (i) a minority level-1 agent of sophistication s lies with probability l(1,s), and (ii) a majority level-1 agent of sophistication s lies with probability l(2,s).

c. Private-Hom: After a contrary signal, a level-1 agent of sophistication s lies with probability l(3,s).

The above proposition implies a particular pattern of lying rates and voting behavior in *private* treatments, as we explain below. If we classify subjects on the basis of scenarios in which they lie, a subject's category will be predictive of his sophistication level. We define four categories, C1-C4, for *Private-Het* and two categories, C5-C6, for *Private-Hom*. Consistent lying at a given information set is defined as lying more than 50% of the time. In the *Private-Het* categories C1 agents lie consistently both in majority and minority, C2 (C3) agents lie consistently only in minority (majority) and C4 agents never lie consistently. In *Private-Hom* categories C5 agents lie consistently in *Private-Hom* while C6 agents do not.

Given the definitions of the behavior of level-k types, C3 behavior is not captured by the model and should be empty. By the law of large numbers, categories C1, C2 and C5 should contain exclusively level-1 agents while categories C4 and C6 concentrate all level-0 subjects as well as some level-1 agents. To see this, recall that a level-1 agent of sophistication s is more likely to lie after a contrary signal if in minority than if in majority. The law of large numbers thus implies that if an agent consistently lies in majority, he must also consistently lie in minority. Consequently, the proposition implies different average sophistication levels across categories C1, C2 and C5. Let $E(s|Cx)$ denote the average sophistication level among Cx-agents. It must be true that

$$E(s|C5) > E(s|C1) > E(s|C2). \tag{2.1}$$

The intuition for the above is as follows. Let threshold s_r, for $r \in \{1, 2, 3\}$, correspond to the s-value at which $l(r, s)$ crosses the horizontal .5 line. Given that $l(1, s) > l(2, s) > l(3, s), \forall s \in [0, 1]$, it is trivially true that $s_3 > s_2 > s_1$. Now, simply note that C5 subjects are defined by $s \geq s_3$ while C1 subjects are defined by $s \geq s_2$ and C2 subjects are defined by $s \in [s_1, s_2)$.

Double inequality (2.1) implies a particular ranking of lying rates across categories. Let $E(l(1, s)|Cx)$ and $E(l(2, s)|Cx)$ denote the average lying rate in respectively minority and majority conditional on being a member of category Cx, for $x < 5$. Similarly, let $E(l(3, s)|Cx)$ denote the average lying rate conditional on being a member of category Cx, for $x \geq 5$. It must be true that $E(l(1, s)|C1) > E(l(1, s)|C2)$. This follows directly from using (2.1) together

with the fact that $l(1,s)$ is increasing in s, for $s \in [0,1]$. Double inequality (2.1) in contrast does not pin down the relative size of $E(l(3,s)|C5)$ and $E(l(2,s)|C1)$. Indeed, two effects oppose each other. On the one hand, C5 subjects are more sophisticated on average than C1 subjects as seen earlier (we call this the *selection effect*). On the other hand, for any given s it holds true that $l(2,s) > l(3,s)$, i.e., lying is more intuitive the fewer agents share one's preference type. When comparing a C5 and a C1 majority subject sharing the same s, the C5 subject's probability of lying after a contrary signal is thus strictly lower than that of the C1 majority subject (we call this the *size effect*). Which of the two effects dominates is a priori unclear, so that we cannot make a clear prediction of the ordering of lying rates for C1 and C5 subjects.

Double inequality (2.1) also implies a particular ranking of voting rates across categories. Let $E(d(s)|Cx)$ denote the average rate of applying $\Lambda(1)$ (as opposed to $\Lambda(0)$) to the observed signal profile conditional on being a member of category Cx. It must be true that

$$E(d(s)|C5) > E(d(s)|C1) > E(d(s)|C2) > E(d(s)|C4). \tag{2.2}$$

This follows from using (2.1) together with the fact that $d(s)$ is assumed increasing in s. We derive the following hypothesis for the cognitive heterogeneity model.[15] This follows from using (2.1) together with the fact that $d(s)$ is assumed increasing in s. We derive the following set of hypotheses for the cognitive heterogeneity model.[16]

Hypothesis 2.4.
a. The lying rate after a contrary signal in minority of C1 subjects is higher than that of C2 subjects.
b. The average frequency of a conform decision given one conform signal of highest for C5 subjects, followed by C1, C2, and C4 subjects. In other words, there is a significant correlation between consistently lying after conform signals and consistently applying $\Lambda(1)$ *to the observed signal profile.*

[15] In addition to the predictions on treatment behavior, one can also derive predictions on the post-experimental tests: a. in the SCT C1 subjects perform better than C2 subjects who themselves perform better than C4 subjects, b. C5 have the highest IDT threshold, followed by C1, C2 and C4 subjects.

[16] In addition to the predictions on treatment behavior, one can also derive predictions on the post-experimental tests: a. in the SCT C1 subjects perform better than C2 subjects who themselves perform better than C4 subjects, b. C5 have the highest IDT threshold, followed by C1, C2 and C4 subjects.

As a last note, we assume that there is a large majority of level-0 subjects among subjects, so that the assumption made by level-1 players is empirically approximately correct. This implies that their lying should be payoff improving.

2.6.2 Results: Disaggregating behavior

In what follows, we analyze whether the predictions of the cognitive heterogeneity model are born out by the data. First, we explore whether there is indeed heterogeneity in behavior that is consistent with our dicotomy of level-0 and level-1 types. Subsequently, we examine the predictive power of the set of hypotheses 4.[17]

Evidence of Heterogeneous Behavior

We analyze (a) whether there is a fraction of subjects who consistently lie in *Private-Het* minority, *Private-Het* majority and *Private-Hom* after a contrary signal, (b) whether lying after a contrary signal is indeed payoff-increasing, and (c) whether a fraction of subjects consistently applies the optimal decision rule $\Lambda(1)$ to the observed signal profile in all treatments.

We start by analyzing behavior at the communication stage in Table 2.5. Columns *all* show the average lying rate of all subjects, just as in Table 2.3. In addition, columns *liars* indicate the average lying rate among subjects who lied at least once at a given information set. The idea is to analyze whether subjects who lie once lie very frequently. Columns *share* indicate the share of subjects categorized as liars at a given information set.

For all information sets and both treatments *Private-Hom* and *Private-Het*, the lying rate after a contrary signal increases strongly when going from the *all* column to the *liars* column, the latter featuring lying rates after contrary signals between 44.5% and 63.0%. The share of liars after a contrary signal is between 22.9% and 33.3%. In all three scenarios, the large majority of subjects thus never lies after a contrary signal while a small share of subjects instead

[17]Please refer to Appendix 2.8.2 and 2.8.2 for the analysis of the predicted ordering in the post-experimental tests. We refer to results when appropriate.

appears to lie often.[18]

We define consistent lying at a given information set as lying at least 50% of the time. We find that 9.4% of *Private-Hom* subjects lie consistently after a contrary signal, 15.21% in *Private-Het* majority and 25% in *Private-Het* minority. The relative size of these three groups supports our assumption that lying is more intuitive (and probable), the higher the number of subjects of the other preference type.

Table 2.5: Heterogeneous lying rates in *private* treatments in %

Signal	Private-Hom			Private-Het		
	all	liars	share	all	liars	share
contrary	10.2	44.5	22.9			
contrary in min.				21.9	63.0	33.3
contrary in maj.				14.9	44.7	33.3
conform	0.7	22.3	3.1			
conform in min.				1.0	29.8	3.1
conform in maj.				0.5	15.9	3.1

Notes: Column *all* indicates the overall lying rate, column *liars* the lying rate of subjects that lie at least once in the corresponding category, and *share* indicates the share of liars for each category.

In the cognitive heterogeneity model lying is payoff-increasing because only a small fraction of cognitively sophisticated subjects (level-1 subjects) seizes the available profitable lying opportunity. Table 2.6 reports results from mixed-effects regressions with profits per period as the dependent variable. Regression (1) includes data from both *private* treatments (1) while (2) only includes data from *Private-Het*.

Regression (1) indicates that lying is generally profitable. On average, a lie increases payoffs significantly from 46.77 to 58.97 tokens. The non-significance of the *Hom* and *Lie Contrary*Hom* coefficients in (1) indicates that the profitability of lying does not depend on the group composition being homogeneous

[18]The lying rates after conform signals also increase in the *liars* column. Note, however, that the share of liars in this category is very small (3.1%).

or heterogeneous. Moreover, the non-significance of the *Minority* and *Lie Contrary*Minority* coefficients in (2) indicates that the profitability of lying does not depend on being in majority or minority.[19]

Table 2.6: Lying and Payoffs

	(1) Private	(2) Private-Het
Lie Contrary	12.20***	12.70**
	(4.32)	(5.38)
Hom	4.52	
	(3.06)	
Lie Contrary*Hom	-4.71	
	(6.72)	
Minority		-0.02
		(3.56)
Lie Contrary*Minority		-2.08
		(8.47)
Constant	46.77***	46.83***
	(2.21)	(2.26)
Obs.	1,978	959
# of Groups	32	16
# of Ind.	192	96

Notes: This table reports coefficients using a linear panel model with mixed effects. Regression (1) includes a dummy *Lie Contrary* equal to 1 if lying after a contrary signal and 0 otherwise, a dummy *Hom* equal to 1 for *Private-Hom* and 0 for *Private-Het*, as well as an interaction term *Lie Contrary*Hom*. Regression (2) controls for being in minority (*minority*) and lying in minority (*Lie Contrary*Minority*). Standard errors in parentheses.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

As to voting behavior, across all treatments a significant proportion of subjects consistently applies $\Lambda(1)$ to the observed signal profile. In *Public-Het*

[19]Individual lying implies a coordination problem. If two subjects of the same preference type and both holding a contrary signal lie simultaneously, the triggered shift in the decision rule will be excessive. We indeed find a decrease in profits when there are two simultaneous lies, but these cases only happen extreme rarely, i.e., in less than 2% of cases, this being a trivial consequence of the low aggregate lying rate and of the small committee size. See Appendix 2.8.2 for an analysis.

and *Public-Hom* we find that a share of respectively 41.66% and 44.79% of subjects consistently (i.e., more than 50% of the time) votes for the conform decision given an observed signal profile containing a unique conform signal. In *Private-Het* and *Private-Hom*, these shares decrease to respectively 29.17% and 28.13%. The decrease in shares when going from *public* to *private* treatments can be explained by the skepticism towards information following from communication.[20]

Lying and voting behavior by categories

Table 2.7 shows for all categories C1-C6 the number of subjects in each category, the lying rates after a contrary signal in minority and majority, and the frequency of a vote for the conform jar given an observed signal profile containing one conform signal. In line with previous results, the categories C4 and C6 contain the vast majority of subjects, 72% in *Private-Het* and 91% in *Private-Hom*. According to the cognitive heterogeneity model, these correspond mostly to level-0 agents. C1 and C2 subjects together constitute roughly 25% of *Private-Het* subjects while C5 subjects constitute 9% of subjects in *Private-Hom*. These three categories of subjects correspond to level-1 agents in the model. Finally, the number of C3 subjects is very low (3%) as predicted.[21]

In minority, the lying rate of C1 subjects (80.2%) is higher than that of C2 subjects (71,5%), as predicted. Recall that the intuition is that category C1 subjects exhibit a higher average level of sophistication than C2 subjects. In addition, we find that the lying rate of the C5 subjects is higher than of C1 subjects (85.6% vs 77.6%), which suggests that the selection effect (i.e., being in C5 requires a higher sophistication level than being in C1) dominates the size effect (i.e., lying as majority type in a heterogeneous committee is more intuitive than lying in a homogeneous committee). Finally, C4 agents have a very low average lying rate in both minority and majority, which is compatible

[20] We find similar shares in the IDT where 27.34% of all subject apply the optimal decision rule and 70.31% apply the majority heuristic. For further information see 2.8.2.

[21] For the following analysis a caveat applies. To analyze the predicted order in lying and voting behavior, we compare empirical frequencies across selected categories of subjects, as opposed to using statistical tests. One key reason is that we disaggregate behavior across subgroups of limited size, which implies low statistical power.

Table 2.7: Lying and voting behavior by categories

	Category	Obs	Lying rates		Vote
			Minority	Majority	Conform
Het	C1	11	80.2%	77.6%	66.9%
	C2	12	71.5%	10.4%	49.5%
	C3	3	6.7%	55.7%	42.6%
	C4	66	3.9%	3.5%	23.5%
Hom	C5	9	85.6%		70.3%
	C6	87	2.4%		25.1%

Notes: In *Private-Het*, we could not categorize 4 subjects because they did not receive a contrary signal in minority.

with these being to a large extent level-0 agents who always truth-tell.[22] We also find the predicted ordering in the voting behavior: C5 vote more often conform than C1, C1 more often than C2 and C2 more often than C4.

Result 2.4.
The lying rate after a contrary signal in minority of C1 subjects is higher than the one of C2 subjects. Similarly, we find that C5 most often vote conform given a signal profile containing one conform signal, followed by C1, C2 and C4 subjects. There is thus a positive correlation between consistently lying after contrary signals and consistently applying $\Lambda(1)$ *to the observed signal profile.*

We find results that are consistent with the main predictions of the cognitive heterogeneity model. At the communication stage in *private* treatments, a small fraction of subjects (17% on average across treatments) consistently lies after contrary signals while the vast majority of subjects always truth-tells. Across treatments, roughly 35% of subjects consistently apply their type-specific payoff-maximizing decision rule. Finally and most importantly, consistent lying after contrary signals is strongly associated with applying the type-specific payoff-maximizing decision rule. We thus identify two groups of agents who correspond roughly to level-1 and -0 agents in the cognitive heterogeneity model.

[22] Appendix 2.8.2 analyzes whether subjects in higher categories also perform better in the SCT. We find no clear ranking of performance in the SCT. Results suggest that subjects did not understand the (quite complex) SCT and therefore continued acting as in the treatment, although optimal behavior in the SCT deviates from optimal behavior in the treatment.

2.7 Conclusion

This paper reports results from a 2x2 experimental design aimed at understanding the drivers of individual behavior in a simple communication and voting game featuring known heterogeneous preference types. Besides the standard model of self-interested and strategic agents, we also tested models of social preferences and of naïve voters. Aggregate behavior is not consistent with any of the models assuming homogenous agents. Further disaggregating results, however, we find heterogeneous individual behavior that is consistent with two cognitive sophistication levels. The numerically dominant naïve subjects do not maximize payoffs but rather truth-tell and predominantly vote with the majority of signals. In contrast, sophisticated subjects do follow their type-specific payoff-maximizing decision rule and lie in a way that allows them to influence the commitee's decision in their favor.

Our experimental findings caution against interpreting low lying rates as homogenous truthful communication (with a low degree of lying). Rather, the lying rates appear to reflect the presence of a small share of sophisticated consistent liars facing a majority of unsophisticated truth-tellers. More broadly, this paper highlights the need to integrate cognitive heterogeneity into theoretical models of committee decision-making and build mechanisms that are resilient to naïve exploitation. The mechanism design literature has a growing body of works that impose deviations from rationality (e.g., no preference maximization (Clippel, 2014), varying but bounded "depths of rationality" (Saran, 2016)). These may provide a starting point for future theoretical work on committee design.

Though this experiment finds no role for social preferences, richer deliberation processes may contradict this conclusion. Debate might in some cases stimulate empathy, solidarity and common identity while it may in other cases reinforce *in vs outgroup* dichotomies and cause preference polarization. Future experiments ought thus to examine other deliberation protocols (e.g., sequential, repeated, subgroup-based).

2.8 Appendix

2.8.1 Post-experimental Tests

The following post-experimental tests were conducted: the strategic communication test (SCT), the individual decision test (IDT), the lying aversion test and the social value orientation slider.

The SCT test evaluates subjects' ability to communicate strategically. It is only taken by subjects in the *private* treatments (as these involve communication) and quasi-replicates the treatment game. A subject keeps his preference type from the treatment. Other subjects are now substituted with computers whose known strategy is to truthfully announce their signals and vote sincerely under the assumption of truth-telling by others. In the SCT a subject only chooses his announcement in the communication stage. At the voting stage, he is replaced by a computer which votes sincerely on the basis of the subject's signal and others' (truthfully announced) signals. Payoffs obtained by the two computerized committee members are randomly allocated to two treatment participants. We use the strategy method to elicit all choices conditional on being in minority or majority and the available signal. In *Het* subjects face four scenarios: one is either in majority or in minority and one either holds a contrary or a conform signal. Of these, only the minority and contrary signal scenario provides a payoff-incentive to lie. In *Hom* subjects face two scenarios. The committee is homogeneous and one holds either a contrary or a conform signal. In both of these cases truth-telling is payoff-maximizing.

The second test is the IDT which evaluates the ability to choose the optimal decision rule. A subject observes three signals as in the *public* treatments but now chooses a jar alone. As compared to the treatments, the IDT excludes effects related to beliefs about others' behavior or social preferences. We use the strategy method. Subjects make a decision for each of the four possible signal profiles, as we seek to identify the minimal number of conform signals required by a subject to choose the conform decision. A subject requiring a minimum of x conform signals to choose the conform decision is said to follow the threshold rule x. On the basis of IDT behavior, we assign threshold rule x to a given subject if the difference between 4 and his total number of conform

decisions is x.[23]

The third test is a lying aversion test based on Gneezy, Rockenbach, and Serra-Garcia (2013). It is a two-player deception game where the sender's decision to lie increases own payment independent of the receiver's decision (see the original paper for more details). In contrast to Gneezy, Rockenbach, and Serra-Garcia (2013), any subject is assigned twice to a two-persons matching group and plays the game once as a sender and once as a receiver. We only use the decision made by subjects when acting as sender. We furthermore only let subjects play the game once in each role. Our test results replicate those of Gneezy, Rockenbach, and Serra-Garcia (2013).

The fourth test is a social value orientation slider aimed at measuring social preferences (Murphy, Ackermann, and Handgraaf, 2011). At the end of the experiment, subjects answered a questionnaire gathering information about their risk aversion, trust of others, and demographic characteristics. Subjects were also asked specific questions on how they played and underlying motives.

2.8.2 Additional Analysis

Lying aversion

Lying behavior in the experiment may have been affected by agents' lying aversion, which was measured in the post-treatment lying aversion test. To test this hypothesis, we run three different regressions. Regression (1) is a discrete choice model with the dummy variable *lie given a contrary signal* as dependent variable. In regressions (2) and (3), we use a linear regression and take as dependent variable the number of lies during the 20 periods. In regression (2) we include all subjects, while in regression (3) we only include subjects who lied at least once. Regressions (1) and (2) allow us to test whether the independent variables influence respectively the probability to lie or the frequency of lying over the 20 rounds. We in addition use the restriction *at least one lie* for regression (3), as we conjecture that subjects who lied at least once were more likely to identify the lying incentive. We use as independent variables the treatment dummy *Het*, *Period* to control for learning effects,

[23]Two caveats are in order. First, the assignment method rests on the assumption that subjects' decision rule is monotonic in the number of conform signals. Second, our method does not allow us to observe whether a subject's decision rule is stochastic as opposed to deterministic.

the dummy variable *SCT* to check for comprehension of lying incentives[24], *IDT treshold*, *lying aversion* and *SVO*. We find that lying aversion exclusively influenced the behavior of those subjects who lied at least once. The variable *lying aversion* has no significant influence in either regression (1) or regression (2). As soon as we drop all non-lying subjects from regression (3), we find that lying aversion negative impacts the number of lies.

Risk attitude and Decision-Making

The post-experimental questionnaire contained a non-incentivized question on risk attitudes. This question was taken from the German SocioEconomic Panel (SOEP). Subjects were asked about their "willingness to take risks in general", and had to indicate their answer on a scale ranging from 0 ("risk averse") to 10 ("fully prepared to take risks"). This measure was found to highly correlate with incentivized measures on risk attitudes (Dohmen et al., 2011). The variable *risk* in the regression below corresponds to this measure.

To test whether risk attitudes influenced behavior, we run a regression for the *public* treatments where the dependent variable is a dummy equal 1 if the subject votes for the conform decision and 0 otherwise. Besides risk attitude (as retrieved from the post-experiment questionnaire), independent variables include subjects' IDT threshold, dummy variables for the number of conform signals and the A-levels math grade. The coefficient for risk aversion is marginally significant ($p \leq 0.10$) and small in size, in contrast to the coefficients for the IDT threshold and the dummies for the number of conform messages. We therefore conclude that risk aversion had a negligible impact on behavior.

Potential Coordination Problem of Lying

As outlined in section 2.6.2, we report payoffs of (majority) types after respectively one and two simultaneous lies in Table 2.10 to analyze how payoffs depend on the number of simultaneous liars in a committee. We identify all *private* treatment aggregate signal realizations in which two subjects of the same preference type hold a contrary signal. These are the instances where

[24] We here use the answer from our first question in the SCT. Recall that it was rational to lie in *Private-Het*, but not in *Private-Hom*. The test is useful as a proxy for comprehension of lying incentives.

Table 2.8: Impact of lying aversion on lying behavior

	Lie	Number of lies	Number of lies
Het	0.05**	0.30	-0.97*
	(0.02)	(0.30)	(0.56)
Period	0.001***		
	(0.000742)		
SCT	0.18***	4.48***	3.42***
	(0.02)	(0.70)	(0.77)
IDT threshold	-0.04*	-0.68	-1.28**
	(0.02)	(0.42)	(0.59)
Lying Aversion	-0.002	-0.03	-0.11**
	(0.002)	(0.02)	(0.04)
SVO	-0.001	-0.02	-0.01
	(0.001)	(0.01)	(0.02)
Constant		2.32**	6.71***
		(0.89)	(1.57)
Obs.	1,978	192	65
# of groups (cluster)	32	32	27
# of individuals	192	192	65

Notes: Regression (1) reports marginal effects calculated at the means of covariates using a logit panel model with mixed effects. The dependent variable is a dummy equal to 1 for a lie after a contrary signal and 0 otherwise. Regression (2) and (3) report coefficients using a linear regression model with standard errors clustered on matching group level. The dependent variable in regression (2) and (3) is the number of lies after a contrary signal over the course of the treatment by a given subject. In regression (3) we restrict the sample to subjects who lied at least once. Independent variables include a dummy *Het* equal to 1 for *Private-Het* and 0 for *Private-Hom*, *Period* taking the value of the corresponding period, the individual scores from the post-experimental tests, *SCT*, *IDT threshold*, *Lying Aversion* and *SVO*. Standard errors in parentheses.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.9: Impact of risk attitude on decision-making

	Public
Het	0.17
	(0.26)
IDT threshold	-1.84***
	(0.26)
1 conform	2.24***
	(0.19)
2 conform	7.50***
	(0.32)
3 conform	9.30***
	(0.64)
Risk	0.10*
	(0.06)
Math Grade	-0.03
	(0.13)
Constant	-0.19
	(0.64)
Observations	3,380
Number of groups	30

Notes: This table reports marginal effects calculated at the means of covariates using a logit panel model with mixed effects. The dependent variable is a dummy equal to 1 for a conform vote and 0 for a contrary vote. Independent variables include a dummy *Het* equal to 1 for *Public-Het* and 0 for *Public-Hom*, the *IDT threshold*, dummy variables for the number of conform signals, *risk* and the A-levels *math grade*.

two majority subjects would each have an incentive to lie unilaterally. We build matching group averages for profits after one lie and after two lies and compare profits. The table indicates that profits as expected decrease when shifting from one to two simultaneous lies. Crucially, however, two simultaneous lies only happen extremely rarely, i.e., in less than 2% of cases. For all practical purposes, a subject lying at the communication stage can thus legitimately assume to be the only one lying, as in the unilateral deviation scenario in the putative truthful-sincere equilibrium analyzed in Coughlan (2000).

Table 2.10: Profits of majority types per number of lies

		1 lie	2 lies
Private-Hom	Average profit	42.75	28.75
	Share of obs.	21.35%	1.69%
Private-Het	Average profit	35.13	20.00
	Share of obs.	29.14%	1.71%

Notes: Only lies by majority types after a contrary signal are counted in cases where two majority types hold a contrary signal.

Strategic Communication Test and Communication in the Treatments

After the *private* treatments subjects took the strategic communication test (SCT) aimed at checking their understanding of strategic lying. If, as argued, lying in the treatment was driven by superior cognitive ability, an intuitive conjecture would be that lying after contrary signals in the treatment correlates positively with better performance in the SCT.

Table 2.11: Lying in SCT in % by categories

	Category	#	SCT lying only min	only maj	min & maj	never
Het	C1	11	9.09	18.18	72.73	0
	C2	12	16.67	25.00	8.33	50.00
	C3	3	33.33	66.67	0	0
	C4	66	0	12.12	6.06	81.82
Hom	C5	9	NA	77.78	NA	22.22
	C6	87	NA	5.75	NA	94.25

Recall that *Private-Het* and *Private-Hom* subjects did not take the exact same SCT. For *Private-Hom*, lying was never individually payoff-improving in the SCT. For *Private-Het* subjects, the only scenario where lying was payoff-improving in the SCT was after a contrary signal in minority. Optimal communication behavior in the SCT thus differed from optimal behavior in the

treatments.

Table 2.11 shows results in the SCT for each of the treatment groups C1-C6. We report lying rates conditional on contrary signals. No clear ranking of performance in the SCT emerges across the considered categories. The only striking regularity is that SCT behavior closely resembles treatment behavior for all categories but C2. For example, most C1 subjects (72.73%) lie both in minority and majority in the SCT, just as in the treatment. Similar insights apply to C3, C4, C5 and C6. Results suggest that subjects did not understand the (quite complex) SCT and simply continued acting as in the treatment (suggesting the presence of order effects), rendering SCT results little informative. In particular, the notion that other subjects were replaced by computers might have caused confusion.

Individual Decision Test and Voting in the Treatments

After the treatments subjects took the individual decision test (IDT) aimed at measuring their decision rule in an individual decision task. The idea is that the IDT gives a cleaner measurement for the decision rule that the treatment as it excludes the role of beliefs and strategic interactions. For the IDT comparison we pool subjects from all treatments since the proportion of individuals applying each IDT threshold does not differ significantly between treatments. In an ordered logistic regression featuring the IDT threshold as the dependent variable, the coefficients of all treatment dummies are insignificant ($p > 0.36$).

Table 2.12: IDT decisions by lying category

	Category	Obs	IDT
Het	C1	11	1.5
	C2	12	1.8
	C3	3	1.67
	C4	66	1.80
Hom	C5	9	1.1
	C6	87	1.8

Notes: In *Private-Het*, we could not categorize 4 subjects because they did not receive a contrary signal in minority.

The IDT also shows the heterogeneity in voting behavior, 0.78% of subjects have an IDT threshold of 0, 27.34% of 1, 70.31% of 2 and 1.56% of 3, thus there are two large groups. More then two thirds of subject apply the majority heuristic and less than one third the optimal decision rule. The IDT threshold also correlates with treatment behavior. Subjects with an IDT threshold of 1 vote conform after one conform signal much more frequently than subjects with an IDT threshold of 2 (63% vs 24%). More importantly, in Table 2.12 we analyze the correlation of lying and IDT behavior. Based on their lying behavior subjects are classified into categories of a presumably higher sophistication which are presumably more likely to use the optimal decision rule in the IDT. As predicted, the IDT threshold characterizing the presumably very sophisticated C5 subjects is very low (1.1) and thus very close to the optimal threshold of 1. C5 subjects' threshold is lower than that of C1 *Private-Het* subjects (1.5). These in turn have a lower threshold than C2, C4 and C6 subjects (1.8). The only deviation from predictions by the cognitive heterogeneity model is the high threshold of C2 subjects (and thus suboptimal) in the light of their good performance in the treatment.

2.8.3 Instructions

We print instructions for the *Public-Hom* blue-biased type (B.1) and for the *Private-Hom* blue-biased type (B.2) treatments. Aspects where the instructions differ for red-biased types are indicated in round brackets. Aspects where the instructions differ for heterogeneous groups are indicated in square brackets. Instructions for post-experimental tests are available upon request.

General explanations for the participants

You are taking part in an economic experiment. Please read the following instructions carefully. You can earn money in this experiment. Your payment will depend on your decisions and on the decisions of the other participants.

During the experiment communication is prohibited. Failure to comply will result in exclusion from the experiment and loss of earnings. Should you have any questions, please address them to us: hold your hand out of the cabin and one of the experimenters will come to your seat.

At the end of the experiment, all sums of money will be paid to you in cash. During the experiment monetary amounts do not correspond to Euro, but to points. In the end, the total point earnings that you obtained during the experiment will be converted into Euro, where: **150 points = 1 Euro.**

The study consists of four parts:

Part 1. Control Questions: you are asked to answer control questions to check comprehension.

Part 2. Experiment: The experiment consists of several parts. Your earning from all parts will be paid.

(1) The instructions for Part 1 can be found below.

(2) You will receive the instructions for the other parts later.

Part 3. End: After the experiment you will receive a questionnaire with general questions. Please fill this out carefully.

Part 4. Payment: You will receive the payment privately. The other participants will not know the amount of your payment.

Instructions Experiment Part 1

Part 1 of the experiment consists of 20 rounds. [At the beginning of the experiment, you will be randomly assigned to a type, type A or type B. The type allocation is maintained throughout the experiment.] In each round, all participants will be divided into groups of 3 participants randomly. [Per group there are either two Type A-participants and a Type B-participant or a Type A-participant and two type B-participants. You will be informed about the group composition at the beginning of each round.] The group allocation is renewed at the beginning of each round. Therefore the group composition changes in each round.

In the experiment you have the task to vote for one of two jars. There are two possible jars, which we call the Red and the Blue Jar. The Red Jar contains 7 red balls and 3 blue balls. The Blue Jar contains 7 blue balls and 3 red balls.

At the beginning of the game one of the two jars will be selected for your group at random. The probability that the Red Jar is selected is 50%. The probability that the

Blue Jar is selected is also 50%. You will not be told which Jar was selected. In Figure 1 you see the Red and the Blue Jar. Figure 2 displays the image of the unknown jar.

Figure 1: Red and Blue Jar

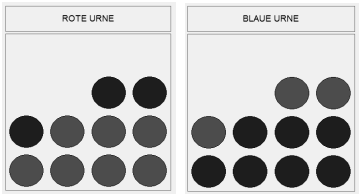
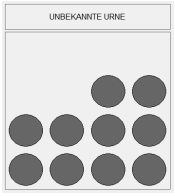


Figure 2: Unknown Jar



As information you will receive the color of three randomly drawn balls from the jar (see Figure 3). In three drawings one ball will be randomly drawn from the jar each time. Each drawing is carried out in two steps:

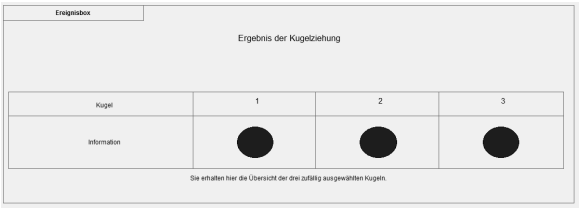
1. A ball is drawn from the jar.
2. The color is written down and the ball is immediately thrown back into the jar.

The number of balls in the jar thus remains the same at each draw. There are three drawings to obtain three balls. Each participant in your group receives the same three balls as information.

Differently colored balls may be drawn from the jar. However, all the balls are drawn from the same jar.

- When the **Red Jar** is selected for your group, each time a ball is drawn from a jar that contains 7 red balls and 3 blue balls.
- When the **Blue Jar** is selected for your group, each time a ball is drawn from a jar that contains 7 blue balls and 3 red balls.

Figure 3: Example for ball draw



After the ball draw the vote takes place. The vote is governed by the following rules:

- If the majority of participants votes for the Red Jar, your group decision is the Red Jar. If there are 2 to 3 votes for the Red Jar and 0 to 1 votes for the Blue Jar, the group decision is therefore the Red Jar.
- If the majority of participants votes for the Blue Jar, your group decision is the Blue Jar. If there are 0 to 1 votes for the Red Jar and 2 to 3 votes for the Blue Jar, the group decision is therefore the Blue Jar.

The payment you receive for the group decision depends on the accuracy of your group decision and of the actual jar.

- If your group decision **corresponds** to the selected Jar and the actual jar is the Red Jar, then you will receive 40 (160) points.
- If your group decision **corresponds** to the selected Jar and the actual jar is the Blue Jar, then you will receive 160 (40) points.
- If your group decision **does not correspond** to the selected jar, then you will receive 10 points.

Table 1: Payments

Number of votes for Red Jar	Number of votes for Blue Jar	Group Decision	Actual Jar	Payment [Type A]	[Payment Type B]
2 or 3	0 or 1	Red Jar	Red Jar	40 (160)	[160]
2 or 3	0 or 1	Red Jar	Blue Jar	10	[10]
0 or 1	2 or 3	Blue Jar	Red Jar	10	[10]
0 or 1	2 or 3	Blue Jar	Blue Jar	160 (40)	[40]

After all participants have voted, the votes will be counted and you will be informed about the outcome of the vote, i.e. votes for Red Jar, votes for Blue Jar, group decision, actual color of the jar and your payment. After the end of the round you will be assigned into new randomly selected groups and the next round begins.

You will receive the payments from all 20 rounds.

If you have questions about the experiment, please contact us now.

General explanations for the participants

You are taking part in an economic experiment. Please read the following instructions carefully. You can earn money in this experiment. Your payment will depend on your decisions and on the decisions of the other participants.

During the experiment communication is prohibited. Failure to comply will result in exclusion from the experiment and loss of earnings. Should you have any questions, please address them to us: hold your hand out of the cabin and one of the experimenters will come to your seat.

At the end of the experiment, all sums of money will be paid to you in cash. During the experiment monetary amounts do not correspond to Euro, but to points. In the end, the total point earnings that you obtained during the experiment will be converted into Euro, where: **150 points = 1 Euro.**

The study consists of four parts:

Part 1. Control Questions: you are asked to answer control questions to check comprehension.

Part 2. Experiment: The experiment consists of several parts. Your earning from all parts will be paid.

(1) The instructions for Part 1 can be found below.

(2) You will receive the instructions for the other parts later.

Part 3. End: After the experiment you will receive a questionnaire with general questions. Please fill this out carefully.

Part 4. Payment: You will receive the payment privately. The other participants will not know the amount of your payment.

Instructions Experiment Part 1

Part 1 of the experiment consists of 20 rounds. [At the beginning of the experiment, you will be randomly assigned to a type, type A or type B. The type allocation is maintained throughout the experiment.] In each round, all participants will be divided into groups of 3 participants randomly. [Per group there are either two Type A-participants and a Type B-participant or a Type A-participant and two type B-participants. You will be informed about the group composition at the beginning of each round.] The group allocation is renewed at the beginning of each round. Therefore the group composition changes in each round.

In the experiment you have the task to vote for one of two jars. There are two possible jars, which we call the Red and the Blue Jar. The Red Jar contains 7 red balls and 3 blue balls. The Blue Jar contains 7 blue balls and 3 red balls.

At the beginning of the game one of the two jars will be selected for your group at random. The probability that the Red Jar is selected is 50%. The probability that the

Blue Jar is selected is also 50%. You will not be told which Jar was selected. In Figure 1 you see the Red and the Blue Jar. Figure 2 displays the image of the unknown jar.

Figure 1: Red and Blue Jar

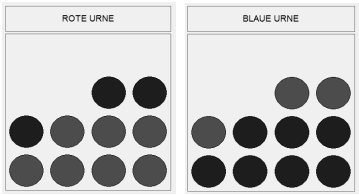
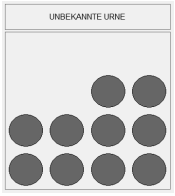


Figure 2: Unknown Jar



As information you will receive the color of three randomly drawn balls from the jar (see Figure 3). In three drawings one ball will be randomly drawn from the jar each time. Each drawing is carried out in two steps:

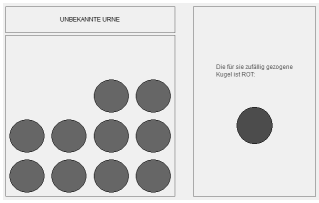
1. A ball is drawn from the jar.
2. The color is written down and the ball is immediately thrown back into the jar.

The number of balls in the jar thus remains the same at each draw.

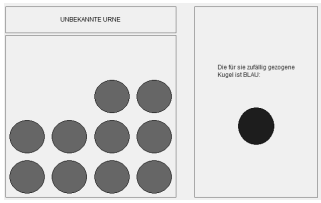
Figure 3 shows that two cases can occur. You either will be shown a red ball (case 1) or a blue ball (case 2).

Figure 3: Example for randomly drawn ball

Case 1) A red ball was drawn



Case 2) A blue ball was drawn



Differently colored balls may be drawn from the jar to the participants of the same group. However, all the balls are drawn from the same jar.

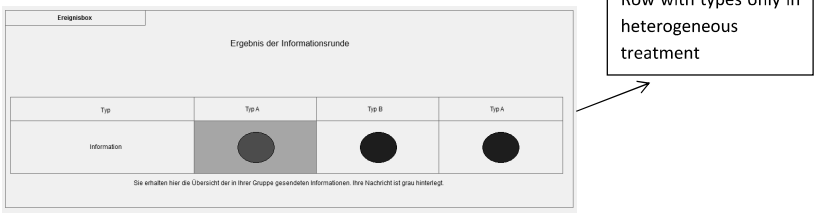
- When the **Red Jar** is selected for your group, each time a ball is drawn from a jar that contains 7 red balls and 3 blue balls.
- When the **Blue Jar** is selected for your group, each time a ball is drawn from a jar that contains 7 blue balls and 3 red balls.

Now there is an information stage. You will send a message about the color of the ball that was shown to you to the other participants in your group. You can choose the content of the message independently of the actual color of the ball (see figure 4).

Figure 4: Message information stage

After you have sent the message, you receive the message of all the other participants of your group (see figure 5). In total, you see 3 messages, the messages of the other two participants and your own message.

Figure 5: Example for results of information stage



After the information stage the vote takes place. The vote is governed by the following rules:

- If the majority of participants votes for the Red Jar, your group decision is the Red Jar. If there are 2 to 3 votes for the Red Jar and 0 to 1 votes for the Blue Jar, the group decision is therefore the Red Jar.
- If the majority of participants votes for the Blue Jar, your group decision is the Blue Jar. If there are 0 to 1 votes for the Red Jar and 2 to 3 votes for the Blue Jar, the group decision is therefore the Blue Jar.

The payment you receive for the group decision depends on the accuracy of your group decision and of the actual jar.

- If your group decision **corresponds** to the selected Jar and the actual jar is the Red Jar, then you will receive 40 (160) points.

- If your group decision **corresponds** to the selected Jar and the actual jar is the Blue Jar, then you will receive 160 (40) points.
- If your group decision **does not correspond** to the selected jar, then you will receive 10 points.

Table 1: Payments

Number of votes for Red Jar	Number of votes for Blue Jar	Group Decision	Actual Jar	Payment [Type A]	[Payment Type B]
2 or 3	0 or 1	Red Jar	Red Jar	40 (160)	[160]
2 or 3	0 or 1	Red Jar	Blue Jar	10	[10]
0 or 1	2 or 3	Blue Jar	Red Jar	10	[10]
0 or 1	2 or 3	Blue Jar	Blue Jar	160 (40)	[40]

After all participants have voted, the votes will be counted and you will be informed about the outcome of the vote, i.e. votes for Red Jar, votes for Blue Jar, group decision, actual color of the jar and your payment. After the end of the round you will be assigned into new randomly selected groups and the next round begins.

You will receive the payments from all 20 rounds.

If you have questions about the experiment, please contact us now.

Chapter 3

Strategic Communication of Endogenous Information and Social Image

3.1 Introduction

Strategic information transmission is a central topic in the economics of information. Following the seminal paper of Crawford and Sobel (1982), a common assumption in this literature is that information is exogenously given to the sender: the information's precision is independent of the sender's characteristics. Yet, in reality, senders often require background knowledge to extract the true informational content of a certain piece of information. Imagine, for instance, a lobbyist who advises the government on regulation policies in the banking industry. When the lobbyist receives new information (e.g., key financial information from banks), she has to possess the necessary skills to adequately interpret the data and extract the true content (e.g., financial risks). The messages that she transmits to the government thus not only convey information about the banking sector, but also about her own expertise. By sending a truthful message the sender can reveal that she was able to draw the correct inferences.[1] When information is endogenous to the sender, as in this example, social image concerns thus may play a role in communication. The

[1] Assuming that at some later stage the receiver will be able to judge the correctness of the message.

sender may enjoy utility from showing others her expertise by truth-telling. This implies that even under misaligned preferences senders who care about how they are perceived by their environment may be induced to report truthfully. Thus, endogenous information combined with social image utility may enlarge the limited scope of truthful information transmission that is typically found in standard cheap talk games (see for cases of inaccurate communication e.g., Austen-Smith (1993), Battaglini (2002), and Coughlan (2000)).

A natural implication is that the scope and the (positive or negative) nature of social image utility depend on the specific type of expertise transmitted by the information. Following the concept of identity utility (Akerlof and Kranton, 2000), different areas of expertise vary in the social status that they convey.

Some knowledge areas may even exhibit a negative social status, such as profound knowledge of drugs, weapons, or tax evasion opportunities. Revealing this type of expertise to other people could induce negative feelings, such as shame. The social status of expertise can be derived from how a given type of knowledge is evaluated in the social system. In organizations, for instance, the value of knowledge is evaluated on the basis of its direct benefits (e.g., the usefulness to achieve a task and its uniqueness or accessibility by others), the manner in which the knowledge was obtained (formal education has a higher perceived value than informal education) and implicit benefits from having the knowledge (feeling pride, power) (for a survey on the value of knowledge see Ford and Staples, 2006).

In this paper, I study endogenous information in a sender-receiver game with misaligned preferences and compare information transmission from two knowledge areas of differing social status. Theoretically, introducing endogenous information into a standard cheap talk game changes the game in such a way that the private information senders receive is not sufficient to infer the state of the world, but requires a certain expertise. For individuals who care about their social image, I show that with increasing social status of information there is a switching point where senders turn from babbling to truthful reporting. Similarly, receivers' probability to trust the message increases with the social status of information.

In the experiment senders receive a multiple choice question with four answer options, of which one option is correct. Senders (receivers) gain a high (low) payoff when the receiver chooses a wrong answer and a low (high) payoff when the receiver chooses the correct answer to the question. In this setting,

a payoff-maximizing sender who does not care about social image is predicted to reveal no information in equilibrium. In the high social status treatment, denoted as *high*, senders receive questions on general knowledge topics (e.g., geography, history, literature), whereas in the low social status treatment, called *low*, questions cover topics such as tabloid TV, celebrities, sports and alcoholic drinks. I find that correct messages in *high* amount to 46%, while only 32% in *low*. Moreover, I show that the driving channel is the ability to signal expertise. When the opportunity to signal expertise is removed, the difference between *high* and *low* vanishes. In addition, I find that that the more difficult the question is, the more likely senders are to report a correct message within the same knowledge area. Also this effect disappears when the possibility to signal expertise through the difficulty of the question is eliminated.

My results demonstrate that endogenous information can enlarge the scope of truth-telling equilibria in sender-receiver games when senders care about being positively perceived by others. This effect of social image replicates findings in individual decision-making environments where people are willing to give up money to signal high ability (Burks et al., 2013; Ewers and Zimmermann, 2015). The present paper also relates to the literature on communication and lying aversion. Compared to standard theory in cheap talk games, which predicts people to misreport information if it is in their material benefit (Crawford and Sobel, 1982), experimental evidence for lying in these games consistently reports overcommunication (e.g., Cai and Wang, 2006; Gneezy, 2005; Sánchez-Pagés and Vorsatz, 2007, 2009). A number of behavioral motives have been recently incorporated into standard models. In a recent meta study on preferences for truth-telling Abeler, Nosenzo, and Raymond (2016) summarize three types of motives: (1) a preference for being honest, i.e., individuals face a lying cost when deviating from the truth (Kartik, 2009; Kartik, Ottaviani, and Squintani, 2007) or gain extra utility from being honest (Ellingsen and Östling, 2010; Sánchez-Pagés and Vorsatz, 2007), (2) a preference for being perceived as honest, i.e., individuals care about their reputation (Fischbacher and Föllmi-Heusi, 2013; Mazar, Amir, and Ariely, 2008; Utikal and Fischbacher, 2013), and (3) social norms, i.e., individuals care about descriptive social norms or social comparisons. For instance, individuals feel less bad about lying when they believe others are lying as well (Diekmann, Przepiorka, and Rauhut, 2015;

Rauhut, 2013).[2] Abeler, Nosenzo, and Raymond (2016) find in their study that combining a preference for being honest with a preference for being seen as honest can best explain the data. This paper adds a further motive for truth-telling: individuals exhibit a preference for showing expertise, something that is crucial once information is endogenous. Thus under endogenous information, reputation takes an even more important role at increasing truth-telling as the reputation for being competent is added to the reputation of being honest.

The remainder of the paper is organized as follows. The next section introduces the model. Section 3 presents the experimental design, Section 4 the results from the experiment, and Section 5 concludes.

3.2 Theoretical Framework

I provide a simple framework that captures the experimental sender-receiver game with endogenous information and illustrates how endogenous information and social image concerns can influence truth-telling.

The game's crucial deviation from a standard sender-receiver game (Crawford and Sobel, 1982) is the endogeneity of information: The information's precision depends on the sender's characteristics. Assume first a standard sender-receiver game. There is a set of two players $N = \{S, R\}$, a sender (S) and a receiver (R). Both sender and receiver know the distribution of the state space. The true state of the world θ is drawn from the state space Θ with uniform distribution over n states. The sender receives a private signal $\widehat{\theta}$ that exclusively contains the true state θ, sends a message $s_S \in \Theta$ to the receiver and the receiver takes an action $s_R \in \Theta$. The receiver believes with probability $\mu(\theta|s_S)$ that the true state is represented by message s_S. Payoffs $\pi_i(\theta, s_R)$ for $i = \{S, R\}$ depend on the receiver's action, not on the message. If the action and the state coincide ($s_R = \theta$), the sender earns a payoff of $\pi_S = 0$ and the

[2] Another strand of literature explains overcommunication by bounded rationality (e.g., quantal response equilibrium and level-k model) (Cai and Wang, 2006). Note that many experimental sender-receiver games involve up to five possible states of the world and a non-linear payoff function In these complex games strategic thinking involving different "depths" is more likely to play a role. Strategic-games with two states of the world and zero-sum structure structure as it is employed in this experiment or individual lying experiments as the widely applied die-rolling paradigm introduced by Fischbacher and Föllmi-Heusi (2013) are considerably easier. In the die-rolling paradigm subjects observe the outcome of a six-sided die roll, report the outcome and receive a payoff proportional to their report.

receiver $\pi_R = 1$; if there is a mismatch ($s_R \neq \theta$), the sender earns a payoff of $\pi_S = 1$ and the receiver $\pi_R = 0$. After the sender has chosen her message and the receiver his action, both players receive their payoff and learn the correct state.[3]

Contrary to the standard game, endogenous information implies that the private signal $\widetilde{\theta}$ the sender receives does not contain the true state of the world, but only some information that together with some background knowledge may allow the sender to extract the true state. Assume a receiver who may face two types of senders that differ in their ability to extract the true state (depending on their background knowledge). With a commonly known probability t the sender is competent and extracts the true state with certainty (high type t_H), and with $1 - t$ she is not competent, cannot extract the true state and plays randomly (low type t_L).[4]

The endogenous information extraction process is thus defined as

$$\widetilde{\theta}(t) = \begin{cases} \theta & \text{if } t = t_H \\ \emptyset & \text{if } t = t_L \end{cases}$$

In case of a low type, there is no longer a strategic game as the sender cannot condition her message upon the true state and therefore cannot optimize. The expected payoff for the sender and the receiver are $(n-1)/n$ and $1/n$, respectively. In contrast, if the sender is of the high type, I obtain a game akin to the standard sender-receiver game with exogenous information where the sender is perfectly informed about the state of the world. Due to the symmetry of the game I can reduce the strategy space of the players in a straightforward way.[5] The sender can either send a truthful message ($s_S = \theta$) or not ($s_S \neq \theta$). The receiver can either follow the message ($s_R = s_S$) or he can decide not to follow ($s_R \neq s_S$). σ_S denotes the sender's mixed strategy to send the correct message and σ_R the receiver's mixed strategy to follow the message.

I assume that utility has two sources: monetary payoffs and image utility. Through communication the sender can show to the receiver that she was able

[3] I refer to the sender in the feminine and the receiver in the masculine.

[4] For simplicity I assume that types are binary. Alternatively, one could image continuous types such that types differ in their confidence of knowing the correct answer.

[5] This simplification of the strategy space has been used by Sánchez-Pagés and Vorsatz (2009).

to extract the true state. The utility function follow the ones proposed by Bénabou and Tirole (2006) and, more closely, Ewers and Zimmermann (2015).[6] In contrast to their specifications, I implement a simplification: Senders accurately know their type, i.e., whether they can extract the true state from the given information. Utility is given by

$$U(\pi, \phi, t, s_S) = \pi(\theta, s_R) + \alpha\beta \; I(s_S, t, \phi)$$

with

$$I(\phi, t, s_S) = \begin{cases} 1 & \text{if } \phi = 1 \text{ and } t = t_H \text{ and } s_S = \theta \\ 0 & \text{else} \end{cases}$$

The first part captures utility over money and the second part incorporates image utility. The payoff $\pi_i(\theta, s_R)$ enters linearly in the utility function and all components are additively separable. $I(\phi, t, s_S)$ is an indicator function taking the value of 1 if the sender has signaling ability ($\phi = 1 \in \{0, 1\}$, i.e., receives endogenous information), knows the correct answer ($t = t_H$), reports it truthfully ($s_S = \theta$) and 0 otherwise. It captures that the sender receives image utility by demonstrating her knowledge to the receiver.[7] In the game, only the sender can derive image utility as the receiver's action does not allow to signal knowledge.

The $\alpha \in [0, \overline{\alpha}]$ specifies the individual weight of image utility. Some senders care more about showing their expertise than others. For the analysis, I assume α to be identical for all senders and commonly known. I consider two benchmarks to derive the main theoretical insights: a standard benchmark with $\alpha = 0$ where senders only care about payoffs, and a social image benchmark with $\alpha = 1$.

The $\beta \in [\underline{\beta}, \overline{\beta}]$ specifies the social status of knowledge. The social status of knowledge may depend on a variety of sources: uniqueness of the information, difficulty to access the information, reputation of the respective knowledge

[6] Ewers and Zimmermann (2015) model an individual-decision making task where subjects report private information about their skills. As in Bénabou and Tirole (2006) individuals value a reputation of being skilled.

[7] Note that I hereby rule out that a lucky guess yields image utility. The idea is that the sender cannot feel proud of transmitting the correct information when it was not caused by her knowledge. I follow here an approach taken in philosophical considerations of knowledge. Gettier (1963) pointed out that any definition of knowledge should rule out lucky guesses. Recent work in experimental philosophy shows that this epistemic intuition is shared by people across different cultures (Machery et al., 2015).

areas, etc. Thus, β is the parameter that will be exogenously manipulated in the experiment.

Plugging the game's payoffs into the above utility function yields the following normal-form game representation:

		Receiver	
		Follow	¬ Follow
Sender	Truth	$(\alpha\beta, 1)$	$((1+\alpha\beta, 0)$
	¬ Truth	$(1, 0)$	$\left(\frac{n-2}{n-1}, \frac{1}{n-1}\right)$

Figure 3.1: Normal form representation of the subgame played by t_H

Note that in the case the sender sends a wrong message and the receiver does not follow the message, there is a chance of $1/(n-1)$ that the correct message is picked by the receiver. Therefore, in this case the expected payoff for the sender and the receiver is $(n-2)/(n-1)$ and $1/(n-1)$ respectively.

Standard Benchmark: Assume $\alpha = 0$. Recall that σ_S denotes the sender's strategy to send the correct message and σ_R the receiver's strategy to follow the message. In the standard benchmark, the unique mixed strategy Nash equilibrium is $\sigma_S^* = \frac{1}{n}$, $\sigma_R^* = \frac{1}{n}$ with the corresponding belief $\mu^*(\theta|s_S) = \frac{1}{n}$. Senders say the truth in $\frac{1}{n}$ of the times and receivers follow the message in $\frac{1}{n}$ of all cases.[8] Note that in this equilibrium it does not matter whether the sender actually knows the correct answer or not since in both cases she sends a correct message with a probability of $\frac{1}{n}$.

Social Image Benchmark: Assume $\alpha = 1$. For β sufficiently large, namely $\beta \geq 1$, there exists a unique Nash equilibrium in pure strategies where senders always send the correct message ($\sigma_S^* = 1$) and the receiver always follows ($\sigma_R^* = 1$). The corresponding belief is $\mu^*(\theta|s_S) = 1$. For intermediate β, namely $0 < \beta < 1$ there is a unique Nash equilibrium in mixed strategies defined by $\sigma_S^* = \frac{1}{n}$, $\sigma_R^* = \frac{1}{n} + \frac{1}{n}\,\beta$ with the corresponding belief $\mu^*(\theta|s_S) = \frac{1}{n}$. Senders say the truth in $\frac{1}{n}$ and the receiver's probability to follow the message

[8]The equilibrium analysis is an application of the proof provided in Sánchez-Pagés and Vorsatz (2007) that derive the equilibrium for a cheap-talk game with exogenous information. See Appendix 3.6.1 for the proofs of the Standard Benchmark and the Social Image Benchmark.

is increasing in β with a probability of at least $\frac{1}{n}$. The social image benchmark illustrates that if individuals put a sufficiently large weight on image ($\alpha = 1$) and the social status of knowledge is sufficiently high ($\beta > 1$) there is a fully information-revealing equilibrium where image gains from sending the correct message outweigh the monetary costs.[9]

Note that the value of β is difficult to interpret in reality. As will be shown in Section 3.3, the treatments of varying social status can however be ordered in terms of their β with $\beta_{high} > \beta_{low}$. Notwithstanding the lacking one-to-one mapping of α and β parameters from theory to experimental design, the theoretical predictions derived from the two benchmark cases allow me to derive comparative statics for the experiment.

Hypothesis 3.1.
In sender-receiver games with endogenous information ($\phi = 1$) and misaligned preferences, an increase in the social status of knowledge β makes senders who care about social image ($\alpha > 0$) more likely to tell the truth and receivers more likely to follow the message and to believe that senders tell the truth.

Note that the above hypothesis captures situations where the information provided by the sender stems (partly) from her own knowledge. Predictions are different in situations where the sender cannot signal knowledge ($\phi = 0$), for instance when information is exogenous. In that case the sender derives no longer image utility. This means that the particular type of knowledge area should not affect the sender's communication behavior.

Hypothesis 3.2.
In sender-receiver games with exogenous information ($\phi = 0$) and misaligned preferences an increase in the social status of knowledge (β) has no effect on the communication of senders, the trusting behavior of receivers and receivers' beliefs about senders' behavior.

3.3 Experimental Design

The theoretical framework suggests that truthful communication of information depends on (a) the ability to signal knowledge (i.e., whether the sender

[9]More generally speaking, an increase in β reduces the minimum value of α such that the condition $\alpha\beta > 1$ for a fully revealing equilibrium is fulfilled.

endogenously acquired the information), and (b) the social image she derives from revealing this information. To test the hypotheses derived in Section 3.2, I conduct a 2x2 between-subjects design where I vary (1) whether the information is endogenous, i.e., the sender is able to signal her knowledge (ϕ), and (2) the social status of the knowledge area (β). On a within-subjects level I vary the difficulty level of questions, i.e., the likelihood that the sender can extract the correct state from the given information. Thus, the social status (β) is varied along two dimensions: the knowledge area and the difficulty.

Table 3.1: Treatments

		Signaling ability	
		Yes	No
Social Status	Low	Signaling-Low	No-Signaling-Low
	High	Signaling-High	No-Signaling-High

In the sender-receiver game senders receive multiple-choice questions from two different knowledge areas. In the *low* social status treatment I use questions from the tabloid press (henceforth denoted gossip questions). Topics include tv-series, music, alcoholic beverages, and commercials. The *high* social status treatment employs questions covering various general knowledge topics such as history, geography, economics, and art (henceforth denoted knowledge questions) (see Table 3.2 for exemplary questions and Appendix 3.6.3 for an overview of all questions and summary statistics).

The questions were tested in a pre-study to measure the difficulty and the social status of the respective knowledge area. After a completely unrelated experiment, 96 subjects received 50 trivia questions ouf of which 48 received knowledge and 48 subjects gossip questions. For each correctly answered question they earned a prize of 6 cent, in total up to 3 Euro.

Out of the 50 questions, 15 were selected for each treatment of the experiment (and a further 15 for a post-experimental test). The 15 questions can be grouped into three levels of difficulty by the frequency of correct answers. There are five easy (80-90%), five intermediate (55-75%) and five hard questions (40-50%). Note that the hardest level of 40% is clearly above the level of 25% that random guessing would produce. The average difficulty level of all 15 questions is nearly the same across treatments (62.9% with standard deviation of 15.2% in *low* vs. 63.1% with standard deviation of 14.5% in *high*). In all treatments, senders and receivers are informed about the difficulty of

Table 3.2: Examplary questions

Treatment	Question and Answers	Correct	Easiness
Low	Who left the boy band Take That in 1995? a) Gary Barlow b) Mark Owen c) Jason Orange d) Robbie Williams	d)	easy (90%)
Low	Which actor plays Bilbo Baggins in The Hobbit? a) Elijah Wood b) Benedict Cumberbatch c) Morgan Freeman d) Martin Freeman	d)	difficult (40 %)
High	Which gemstone is green? a) Opal b) Ruby c) Emerald d) Sapphire	c)	easy (90%)
High	What is the name of the mathematician credited with a famous concept in game theory, named after him? a) Carl Friedrich Gauss b) Alan Turing c) Bernard Bolzano d) John Nash	d)	difficult (40 %)

Notes: The number in brackets in column *Easiness* denotes the percentage of correct answers in the pre-study.

each question: they receive the percentage of correct answers for each question (rounded to the 5%-level). Beliefs about the difficulty of the answer are thus held constant across treatments as well as sender and receivers.

To verify that the questions in *high* and *low* actually evoked a different social status, I included a social status elicitation in the pre-study. Subjects were asked to assess how being good at answering the questions correlates with a set of six characteristics (intelligence, memory, success in studies and life, curiosity, openness for experiences and extraversion) and to indicate how important they perceive these characteristics.[10] To elicit a *social* image and not subjective assessments for the first question, subjects were told to choose the answer they thought was chosen by most participants (for a similar procedure see Krupka and Weber (2013)). I find that the questions from the two knowledge areas clearly created a different social status. In 4 out of 6 characteristics (curiosity, success, memory and IQ) the *high* questions were significantly evaluated higher than the *low* questions. Extraversion was significantly more highly evaluated in *low* and for openness there are no differences between treatments. I also find significant differences between both knowledge areas when I construct an weighted average over all characteristics for each

[10]Furnham and Chamorro-Premuzic (2006) measure the correlation between general knowledge, personality and intelligence. Their personality measures included some of the above-mentioned characteristics. In their experiments they found a consistent positive correlation of general knowledge with general intelligence (Wonderlic test for fluid and crystallized intelligence) and a positive, albeit less consistent, correlation with openness.

subject that weights each characteristic by its perceived importance. This so-called social status score amounts to 0.61 in *high* and 0.27 in *low* (two-sided Mann-Whitney Test, $p = 0.0048$, MW henceforth).[11]

The sender-receiver game with questions from the respective knowledge areas is implemented (a) with signaling ability and (b) without signaling ability, i.e., whether the sender is able to signal her expertise via the message. In the *signaling* treatments senders can extract the true state by solving the multiple-choice question. In the *no-signaling* treatments senders only receive the question (and not the four answer options) and a randomly picked answer from the pre-study. The senders learn only the letter of the picked answer (a, b, c or d) such that they cannot make any inference about the correctness of the random answer. Nonetheless, the subjects know the likelihood of the randomly picked answer to be correct. The instructions make clear that the probability of the randomly picked answer from the pre-study to be correct is equal to the percentage of correct answers. Consequently, the precision of the information does not depend on the sender's characteristics but is exogenous. In the *no-signaling* treatments the sender's choice to truthfully reveal the pre-study's signal can therefore not be driven by image concerns.

Experimental Procedure The experiment contains six parts: (1) sender-receiver game, (2) belief elicitation, (3) social status elicitation, (4) an expertise task, (5) a socio-demographics questionnaire, and (6) revelation of results from (1)-(4).

In the sender-receiver game subjects were randomly assigned to one of two roles, participant A (sender) or participant B (receiver). The game is played 15 times as a one-shot game without feedback, i.e., first all senders complete their task, they choose a message for each of the 15 questions. Then the receivers take an action choice for each question and its corresponding message. The sequence in which questions occur is randomly determined for each session. For each question senders and receivers are randomly matched.[12] The four answers to the questions are labeled with a, b, c, or d and the message the receiver obtains only contains the classifier a, b, c, or d. The receiver knows the question but not the answer options, such that he cannot infer the correct

[11] For further details on the elicitation procedure, results and the social status measure see Appendix 3.6.2.

[12] Given that the game is one-shot and players receive no feedback, the matching procedure should play no role.

answer. Payoffs are as follows. If the receiver chooses the correct answer to the question, he earns 9 Euro and the sender 6; when an incorrect answer is chosen, the receiver earns 6 Euro and the sender 9 Euro. At the end of the experiment one of the questions is randomly chosen for payment.

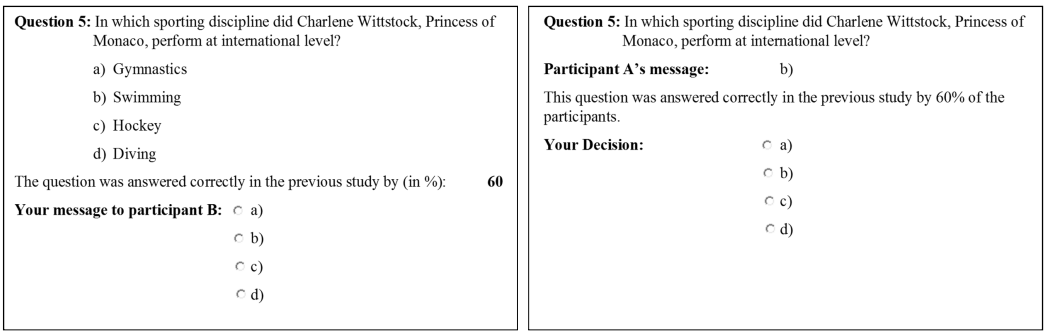


Figure 3.2: Decision of sender and receiver in *signaling* treatments

Notes: The left figure shows the sender's decision and the right figure the receiver's decision. In the *no-signaling* treatments the sender's decision screen does not contain the four answer options, but the answer from the randomly selected participant from the pre-study. The receiver's screen is the same in all treatments.

After the choices (but before revelation of results) beliefs are elicited. Both senders and receivers are asked to indicate their belief about senders' and receivers' behavior. In the *signaling* treatments, subjects report their guess about the average number of correct messages and the average number of times receivers followed the message (out of the 15 questions averaged over all senders respectively receivers). In the *no-signaling* treatments the corresponding belief question about senders' behavior is to guess the average number of messages equal to the random answer from the pre-study. In the *signaling* treatments, senders were additionally asked to voluntarily indicate how many correct answers (a) they knew and (b) they believe the other senders knew (both not-incentivized). The belief elicitation was not previously announced such that prior choices were not influenced. One of the two belief questions was randomly picked for payment. Subjects were paid 2 Euro in case their guess was equal or +/- 1 to the actual number and 2 Euros divided by the absolute estimation error if the estimate deviated by more than 1.

In part 3 I elicited the social status of information similar as in the pre-

study.[13] In part 4, the expertise task, subjects answered 15 further questions of the same knowledge area as in their corresponding sender-receiver game (see for an overview of questions used Appendix Table 3.7 for *high* and Table 3.6 for *low*). Subjects were incentivized to answer correctly the questions and were paid 25 cents for each correct answer. In part 5 subjects answered a socio-demographics questionnaire that included questions on age, gender, nationality and school grades. At the very end of the experiments subjects were informed about the results. In the *signaling* treatments they learned the correct answer, the message, the receiver's action and their payoffs for each question. In the *signaling* treatments they learned the correct answer, the pre-study's answer, the message, the receiver's action and their payoffs for each question. The feedback was provided in a table. The questions and the answer options were not shown, but only the question numbers and labels (a, b, c, d for the correct answer, pre-study's answer and message).

I conducted a total of eight sessions in June and July 2015 and four pre-study sessions in June 2015. In the main experiment a total of 188 subjects participated, out of which 46 subjects participated in *signal-gossip*, 48 in *signal-knowledge*, 48 in *no-signal-gossip* and 46 in *no-signal-knowledge*. A session took on average 50 minutes and subjects earned 12.23 euros on average. In the pre-study, 96 subjects participated (48 in each question treatment) and their payment for the pre-study was on average 2.02 euros (in addition to the money they earned in the previous experiment). All sessions were conducted at the BonnEconLab, subjects were recruited via hroot (Bock, Baetge, and Nicklisch, 2014) and the experiment was run using the experimental software z-Tree (Fischbacher, 2007).

[13]The procedure of the social status eliction was slightly adapted (see Appendix 3.6.5). The evaluation of senders in the *signaling treatments* is qualitatively similar to the pre-study. The three characteristics that had significant differences at the highest significance level ($p < 0.01$) in the pre-study, extraversion, success and IQ, are qualitatively equivalent, but fail to reach significance. Aggregate measures as the average over all characteristics or a measure that controls for individual importance also rank *high* questions higher than the *low* questions, but equally fail to reach significance. Note that in the pre-study the prior task was non-strategic and more simple (earn money for correct answer). The strategic thinking induced in the experiment as well as a desire to maintain a positive self-image after lying may have reduced the treatment effect.

3.4 Results

This section first reports results concerning the senders' behavior and receiver's behavior across the two treatment dimensions, the social status of knowledge area and signaling ability. Second, it shows results on the effect of difficulty.

Given the endogeneity of information in the *signaling* treatments any analysis of truthful communication needs to take into account that subjects may have failed to acquire the correct answer to the question. This means that not all incorrect messages are necessarily intentional lies, but some are mistakes. Equivalently, correct messages contain intentional truthful messages of senders as well as unintentionally correct messages of senders who do not know the answer and send a correct message by chance. But since the difficulty level is identical across knowledge areas, I can nonetheless directly compare the share of correct messages between *low* and *high* in the *signaling* treatment.[14]

As information is exogenous in the *no-signaling* treatments, there is no need to take into account mistakes; messages directly reflect intentions. Senders decide whether they want to transmit ("follow") the answer from the pre-study or not. For a "follower", the probability to send a correct message is equal to the probability of receiving a correct answer, the share of correct answers, $P(C)$, in the pre-study. The probability of a non-follower to send a correct message is equal to $1/3\ (1 - P(C))$. If the sender receives an incorrect answer from the pre-study and does not follow the message, she has a chance of $1/3$ to unintentionally send a correct message. I apply this transformation to the binary communication decisions in *no-signaling* treatments in order to compare the share of correct messages across all four treatments. Note that in both *signaling* and *no-signaling* treatments subjects have the same available technology to transmit a correct message on average. In *signaling*, senders can find out the correct answer to the question and send a truthful message; in *no-signaling*, senders can pass over the answer from the pre-study. Under truthful behavior, the average rate of correct messages should be the same in all treatments.

Figure 3.3 depicts the share of correct messages across all four treatments. The most striking feature is that the share of correct messages in *signaling-high* is 13% higher than the second-highest treatment. Comparing the *signaling*

[14]This assumes that the average share of subjects who knows the correct answer is constant across pre-study and experiment or, weaker, that changes between pre-study and experiment are equivalent for both knowledge areas.

treatments, subjects tell significantly more often the correct message when they transmit information on a general knowledge question compared to a gossip question (MW, $p = 0.02$,). Note that in both treatments the lying incentive is effective: the share of correct messages is significantly lower than the share of correct answers from the pre-study (MW, for both treatments $p < 0.01$). In *high* the rate is, nonetheless, above the 25% that a babbling strategy would produce (Wilcoxon signed-rank test, $p < 0.01$, WX henceforth), while the same is not true for *low* (WX, $p = 0.16$). The results suggest that only the general knowledge questions have a sufficiently high positive social status that provides social image and induces truth-telling.

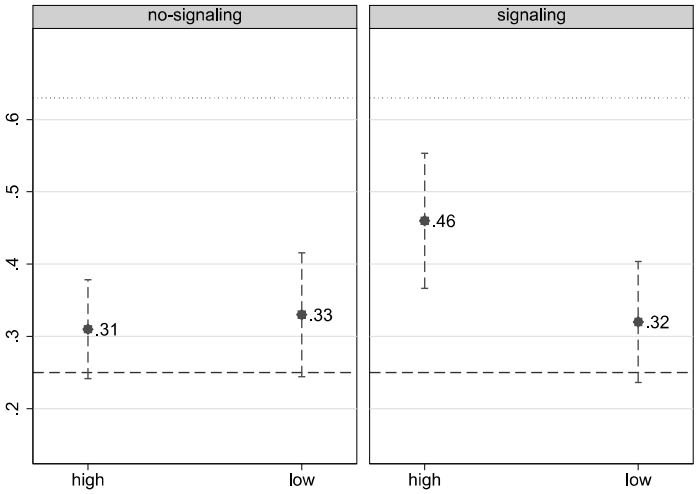


Figure 3.3: Share of correct messages by treatments

Notes: A message is defined as correct if it is equal to the correct answer to the question. The vertical dashed red lines depict 95% confidence intervals. The horizontal dashed black line at 0.25 indicates the predicted rate of correct messages in a babbling equilibrium (benchmark case). The horizontal dotted black line at 0.63 depicts the share of correct answers in the pre-study.

In the *no-signaling* treatments I do not find any significant difference between *low* and *high* (MW, p=0.92). This reveals that endogenous information is crucial for the treatment difference to emerge: Transmitting correctly endogenously acquired information allows senders to signal expertise and derive positive image utility whenever the transmitted information is positively perceived. The fact that the share of correct messages in *signaling-high* is signifi-

cantly higher than in *no-signaling-high* (MW, p=0.02) is further evidence that the possibility to signal expertise in this area induces senders to tell the truth. For the low status knowledge area, *low*, I do not find a statistically significant difference in messages (MW, $p = 0.96$), which reveals that gossip questions do not induce senders to tell the truth even when they can signal expertise in this knowledge area.[15]

Result 3.1.
Subjects transmitting information on general knowledge questions send significantly more often correct messages compared to senders who receive questions on tabloid topics. With exogenous information, the treatment difference in communication disappears.

Turning to the belief data in the *signaling* treatments, I find that senders mildly anticipated the treatment effect, albeit it fails statistical significance. In each treatment, the senders' beliefs about the average number of correct messages relates closely to the actual communication behavior (WX, $p > 0.3$ for both treatments). Yet, senders underestimate the difference in choices between treatments. I find no significant difference in beliefs across knowledge areas (MW, $p = 0.22$). This suggests that senders believe other senders to be on average less influenced by the social status of the knowledge area. On the receivers side the treatment effect was not anticipated (treatment comparison of receivers' belief in *low* vs. *high*, MW, $p > 0.39$). In *high* receivers hold a belief of 31.94% which is less pessimistic than the one of senders at marginal significance (MW, $p = 0.09$). In *low*, receivers' beliefs of 39.13% matches the average belief of senders (MW, $p = 0.36$). The missing anticipation of receivers may be caused by the fact that receivers saw the questions without the four answer options. Although the question itself reveals the knowledge area, the four answer options provide further details. In the *no-signaling treatments* beliefs do not differ across the treatments, which is in line with the actual

[15]Given that the communication behavior is statistically not different from babbling in *no-signaling low* (WX, $p > 0.2$), it is not possible to detect a negative effect of social image, i.e., senders cannot increase their lying behavior beyond babbling. To examine the effect of negatively perceived information, one could use a setup with a lower misalignment of preferences that predicts partial pooling.

behavior of senders.[16] In addition, senders and receivers hold similar beliefs about truth-telling rates.

How do receivers respond to the messages? Trust rates indicate the frequency of receivers to follow the senders' messages. In the *signaling* treatments, receivers in *high* trust in 36.31% of the messages, while in *low* up to 48.41%. The difference is not statistically significant (MW, $p = 0.39$).[17] In both treatments, trust rates are higher than in the benchmark babbling equilibrium (WX, $p = 0.08$ for *high* and $p = 0.004$ for *low*). Neither senders' nor receivers' beliefs about the average trust rate differ across treatment,[18] which is in line with their beliefs on senders' behavior where they neither anticipated the treatment effect. In the *no-signaling* treatments, trust rates are slightly higher, 51.30% in *high* and 55.60% in *low*. However, these differences are not statistically significant (neither within *no-signaling* treatments nor between *signaling* and *no-signaling* treatments). The belief and the trust data thus reveal that the strong treatment difference in the communication of senders in the *signaling* treatments was not anticipated. Consequently, it can be excluded that the treatment effect is driven by strategic considerations, e.g., senders being more honest because they expect receivers being more likely to mistrust them (as it was previously observed in Sutter (2009) or Vanberg (2016)), but senders willingly give up money to signal their expertise.

Result 3.2.
The treatment difference in the truth-telling rates between transmitting endogenous information on general knowledge questions and tabloid topics was not anticipated by senders nor receivers.

The previous results have shown that senders reacted on the higher social status of general knowledge questions and were more inclined to send correct messages. In the following, I analyze whether the difficulty of questions evoked an equivalent effect. Figure 3.4 plots the share of correct messages in the

[16] In the *no-signaling treatments* treatments subjects were asked to indicate the average share of messages equal to the random answer from the pre-study. I therefore apply the same transformation as to the binary communication decisions. The belief about the average share of correct messages is equal to the belief about the average share of follower · average probability of correct answer from pre-study + (1−belief about average share of follower)·(1−average probability of correct answer from pre-study)·1/3.

[17] A random effects regression also finds no statistically significant difference between treatments.

[18] For an overview of all truth-telling and trusting rates (actual behavior and beliefs) see Table 3.8 for *signaling* and 3.9 for *no-signaling* treatments in Appendix.

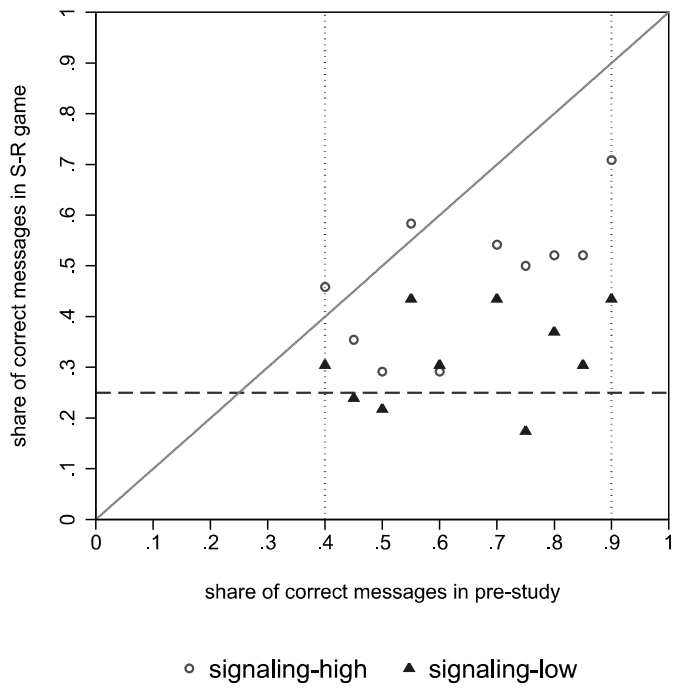


Figure 3.4: Share of correct messages by degree of difficulty in *signaling*

Notes: The solid 45° degree line depicts the share of correct messages in the pre-study. The horizontal dashed line at 0.25 indicates the predicted rate of correct messages in a babbling equilibrium (standard benchmark). The dotted vertical lines at 0.4 and 0.9 indicate the boundaries of the difficulty interval which was used in the S-R game.

sender-receiver (S-R) game in both *signaling* treatments conditional on the degree of difficulty (i.e., the share of correct messages in pre-study). Going from left to right on the x-axis means that the questions become easier and subjects have a higher a priori probability of finding the correct answer. The figure shows that in both treatments the dots indicating the actual share of correct messages lie clearly below the solid 45° degree line, which indicates that senders send intentionally wrong answers. The increasing spread between the solid line and the dots may indicate that the more likely are senders to know the answer, the more likely are they to misreport it. Note, however, that not only an increasing, but also a constant lying rate (independent of the difficulty level) would produce this increasing spread.

To analyze the relationship of difficulty and truth-telling thoroughly, it is therefore useful to impose a statistically testable relationship. I assume that the probability of truth-telling in the experiment $P(T)$ is first and foremost a function of the easiness level of the question, the share of correctly answered questions in the pre-study $P(C)$. Senders can only make a deliberate communication choice when they know the answer. If senders do not know the answer, they cannot make a strategic choice and are forced to send a random answer (as discussed in Section 3.2). Consequently, I assume that all other variables that presumably affect truth-telling behavior depend on the probability of knowing the correct answer (see Equation 3.1). The variables inside the parentheses capture these factors. β_0 is a constant, β_1 measures the effect of easiness ($P(C)$), β_2 the treatment effect of *low*, β_3 the treatment effect of *signaling* and β_4 the interaction effect of *low* and *signaling*. Note that $P(C)$ thus enters the relationship twice, $P(C)$ outside parentheses controls for the effect of knowing the answer and $P(C)$ inside parentheses measures the social status effect of easiness, i.e., a correct answer to a more difficult questions may signal more expertise. For the regression analysis, I expand the function (see Equation 3.2). The regression includes random effects for subjects ϵ_i. The residual is ϵ_{it}.

$$\begin{aligned} P(T) = P(C)\Big(&\beta_0 + \beta_1\ P(C) + \beta_2\ gossip + \beta_3\ signaling \\ &+ \beta_4\ signaling\ gossip\Big) + \epsilon_i + \epsilon_{it} \end{aligned} \tag{3.1}$$

$$\begin{aligned} P(T) = &\beta_0\ P(C) + \beta_1\ P(C)^2 + \beta_2\ gossip\ P(C) + \beta_3\ signaling\ P(C) \\ &+ \beta_4\ signaling\ gossip\ P(C) + \epsilon_i + \epsilon_{it} \end{aligned} \tag{3.2}$$

Table 3.3: Determinants of communication

	(1)	(2)
$P(C)$	0.73***	0.61***
	(0.11)	(0.15)
$P(C)^2$	-0.31***	-0.16
	(0.11)	(0.16)
$low \cdot P(C)$	-0.03	-0.03
	(0.08)	(0.08)
$signaling \cdot P(C)$	0.17**	0.40**
	(0.08)	(0.20)
$signaling \cdot low \cdot P(C)$	-0.20*	-0.20*
	(0.11)	(0.11)
$signaling \cdot P(C)^2$		-0.27
		(0.22)
Observations	1,301	1,301
Number of Subjects	94	94

Notes: This table reports coefficient of a linear random effects regression. $P(C)$ indicates the easiness of the question, i.e., the share of correct answers in the pre-study. All following variables are interacted with $P(C)$. The easiness level itself, $P(C)^2$, the treatment dummy *low* (taking the value of 1 for *low* and 0 otherwise), $low \cdot P(C)$, the treatment dummy *signaling* (taking the value of 1 for *signaling* and 0 otherwise), $signaling \cdot P(C)$, the interaction of both treatment dummies, $signaling \cdot low \cdot P(C)$, and the interaction of easiness with the treatment dummy *signaling*, $signaling \cdot P(C)^2$. Standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.3 reports two random effects estimations. Regression (1) includes all the variables listed above and Regression (2) additionally controls for the interaction of the *signaling* treatments and the difficulty level ($signal \cdot P(C)^2$).[19] In both regressions, the mechanical effect of the difficulty level $P(C)$ is present, i.e., the easier the questions, the more likely senders know the correct answer and tell the truth. The negative coefficient of the squared term $P(C)^2$ in Regression (1) indicates that the rate of truth-telling is increasing in the difficulty of question. Regression (2) controls whether this effect depends on the signaling opportunity. Indeed, Regression (2) shows that this effect is not present in the *no-signaling* treatments, as the coefficient for $P(C)^2$ becomes insignificant. This is intuitive as with exogenously given information senders cannot use the difficulty of the question to signal expertise. As expected, the difficulty of questions matters in the *signaling* treatments: The sum of the coefficients $P(C)^2$ and $signal \cdot P(C)^2$ is statistically different from zero ($p < 0.01$). With endogenous information, senders care more about showing the receiver that they found the correct answer the more difficult a question is. Furthermore, the parametric analysis confirms all non-parametric results (see Regression 1). In the *signaling* treatments, senders in *low* are less likely to tell the truth than in *high* (sum of $low \cdot P(C)$ and $signal \cdot low \cdot P(C)$ coefficients -0.23, $p < 0.01$). The knowledge area, as previously shown, does not matter for the *no-signaling* treatments (-0.03, $p > 0.1$). When senders transmit information on general knowledge questions, senders are more likely to tell the truth when information is endogenous compared to being exogenous (0.17, $p < 0.05$). This positive effect of endogenous information is not present for gossip questions (sum of $signal \cdot P(C)$ and $signal \cdot low \cdot P(C)$ coefficients -0.03, $p > 0.1$).

Result 3.3.
Subjects are more likely to report correct messages when transmitting endogenously acquired information from difficult compared to easy questions. With exogenous information the difficulty effect disappears.

[19] Additional regressions include the score from the expertise test. Expertise is not significantly correlated with the number of correct messages in any of the *signaling* treatments. There is thus no indication for selection effects, i.e., individuals educated in general knowledge to be more likely to tell the truth than individuals competent in gossip questions. Results are available upon request.

3.5 Conclusion

This paper reports evidence on the effect of endogenous information in a sender-receiver game with misaligned interests. Information is considered endogenous if its precision depends on the sender's characteristics, i.e., the sender's ability to extract the true state out of the given information. I demonstrate that senders endowed with endogenous information of high social status tell the truth significantly more often compared to senders in a treatment with low social status information. While in the low status treatment roughly half of all senders who know the correct answer report it truthfully, this share increases to more than 70% in the high status treatment. When senders receive exogenous information, the treatment effect between knowledge areas disappears. Thus, the relevant driving factor is the endogeneity of information which gives senders the possibility to signal their expertise to the receiver.

In the experiment, I control for various factors, which outside of the laboratory would be difficult to achieve: (1) I use questions of the same difficulty level in both knowledge areas and reveal the difficulty level to senders as well as to receivers. In contrast, in reality both factors, social status and difficulty, are likely to correlate, or to be perceived to correlate (e.g., gossip questions may be perceived to be easier than general knowledge questions). (2) In the experiment, the senders' communication has only consequences within the sender-receiver game; there is no room for reputation. In reality, however, many social interactions are repeated and non-anonymous. Thus, developing a reputation for being competent is valuable per se.[20] In particular, expertise in gossip and knowledge question not only differs with regard to its social status but also most likely in the long-term reputational benefits. Revealing to others expertise in general knowledge may, for instance, increase one's reputation in a social network and thus one's career chances. In contrast, tabloid knowledge has little value for academic and most professional purposes. It might even signal self-control problems (e.g., reading online news instead of studying). Consequently, revealing expertise in high status areas compared to low status areas is presumably even more beneficial in reality than in the laboratory.

While this paper finds only positive effects of endogenous information by inducing more senders to report truthfully in the high social status treatment, it is important to note that endogenous information may potentially also de-

[20] Abeler, Nosenzo, and Raymond (2016) make a similar argument for the desire to be perceived as honest.

crease truthful reporting (when individuals have an incentive to communicate truthfully). It can be expected that information which signals some type of misconduct or socially disapproved behavior would produce such a negative effect. For instance, a lobbyist may refrain from reporting tax evasion opportunities to a government, as this information could evoke the image that she herself evades taxes.

More broadly, this paper highlights the need to integrate the endogeneity of information into theoretical models of communication. While I provide a simple model where senders are either able to extract the true state from the private information with certainty or not, it would be instructive to model the information extraction process in further depth. For instance, the sender's accuracy level could be modeled as a continuous variable reflecting that people are often uncertain about their knowledge. Furthermore, the decision to extract the relevant information could be subject to strategic behavior if the information extraction process is costly (i.e., it requires positive effort costs as in solving an intelligence task or analyzing data). Individuals may be inclined to put no effort in solving the task if expected image gains are not large enough to outweigh the effort costs. In this experiment a task with (nearly) no effort cost was chosen deliberately to exclude this effect. Further research may, however, investigate communication of costly endogenous information. This line of research could build upon models of endogenous information acquisition where information is independent of sender's characteristics, but senders make a strategic choice about acquiring the information (Argenziano, Severinov, and Squintani, 2016; Austen-Smith, 1994; Pei, 2015).

3.6 Appendix

3.6.1 Proofs

These proofs are an application of the proof provided in Sánchez-Pagés and Vorsatz (2007), who consider a sender-receiver game with exogenous information and two states of the world. Contrary to Sánchez-Pagés and Vorsatz (2007), this application considers a simplification of the strategy space (*truth*, *follow*) and a utility function with social image concerns.

Standard Benchmark

σ_S (σ_R, respectively) denotes the probability that the sender tells the truth (receivers follows the message).

Proof. Suppose that $0 \leq \sigma_S \leq 1$. I derive the best response correspondence for the receiver who takes the strategy σ_S as given. Suppose that the receiver sends message s_S. In the simplified strategy space, the probability $\mu(\theta|s_S)$ (i.e., the correct answer is given by message s_S) trivially boils down to

$$\mu = \frac{\sigma_S}{\sigma_S + 1 - \sigma_S} = \sigma_S. \tag{3.3}$$

Given μ, the receiver chooses σ_R in order to

$$\max_{\sigma_R} \left(\mu \left(1 \ \sigma_R + 0 \ \ (1 - \sigma_R)\right) + (1 - \mu) \left(0 \ \sigma_R + \frac{1}{n-1} 1 \ \ (1 - \sigma_R)\right)\right), \tag{3.4}$$

which is equivalent to

$$\max_{\sigma_R} \left(\sigma_R \left(\frac{n \ \mu}{n-1} + \frac{1}{n-1} \right) + \frac{1-\mu}{3} \right). \tag{3.5}$$

Therefore, the best correspondence for the receiver is

$$\sigma_R^*(\mu) = \begin{cases} 0 & \text{if } \mu < \frac{1}{n} \\ [0,1] & \text{if } \mu = \frac{1}{n} \\ 1 & \text{if } \mu > \frac{1}{n} \end{cases} \quad \text{or} \quad \sigma_R^*(\sigma_S) = \begin{cases} 0 & \text{if } \sigma_S < \frac{1}{n} \\ [0,1] & \text{if } \sigma_S = \frac{1}{n} \\ 1 & \text{if } \sigma_S > \frac{1}{n} \end{cases} \tag{3.6}$$

Next, I calculate the optimal mixed strategy σ_S^* for the sender. To do so, I consider three cases:

Case A. Suppose that $\sigma_S^* < \frac{1}{n}$. Then, it follows from the optimal behavior of the receiver that $\sigma_R^*(\sigma_S^*) = 0$. Thus, the optimal strategy σ_S^* must be the solution of the following maximization problem:

$$\max_{\sigma_S} \left(1\ \sigma_S + \frac{n-2}{n-1}\ 1\ \ (1-\sigma_S)\right). \tag{3.7}$$

But the solution to his problem is such that $\sigma_S = 1$, which is a contradiction. Hence there does not exist any equilibrium in which $\sigma_S^* < \frac{1}{n}$.

Case B. Suppose that $\sigma_S^* > \frac{1}{n}$. Then, it follows from the optimal behavior of the receiver that $\sigma_R^*(\sigma_S^*) = 1$. Thus, the optimal strategy σ_S^* must be the solution of the following maximization problem: Choose σ_S in order to

$$\max_{\sigma} \left(0\ \sigma_S + 1\ \ (1-\sigma_S)\right). \tag{3.8}$$

But the solution to this problem is such that $\sigma_S = 0$, which is a contradiction. Hence there does not exist any equilibrium in which $\sigma_S^* > \frac{1}{n}$.

Case C. Suppose that $\sigma_S^* = \frac{1}{n}$. Then, it follows from the best response correspondence of the receiver that $\sigma_R^* \in [0,1]$. Thus, the sender faces the problem

$$\max_{\sigma_S} \left(\sigma_S\ \ (0\ \sigma_R + 1\ \ (1-\sigma_R)) + (1-\sigma_S)\left(1\ \sigma_R + \frac{n-2}{n-1}\ 1\ \ (1-\sigma_R)\right)\right), \tag{3.9}$$

which is equivalent to

$$\max_{\sigma_S} \left(\sigma_S \left(\frac{1}{n-1} - \frac{n}{n-1}\sigma_R\right) + \frac{1}{n-1}\ \sigma_R + \frac{n-2}{n-1}\right). \tag{3.10}$$

Hence, the best response correspondence for the sender is

$$\sigma_S^*(\mu) = \begin{cases} 0 & \text{if } \sigma_R < \frac{1}{n} \\ [0,1] & \text{if } \sigma_R = \frac{1}{n} \\ 1 & \text{if } \sigma_R > \frac{1}{n} \end{cases} \tag{3.11}$$

From inspection we see that the mixed strategy $(\sigma_S^*, \sigma_R^*) = (\frac{1}{n}, \frac{1}{n})$ with belief $\mu^* = \frac{1}{n}$ can be sustained as equilibrium strategy.

□

Social Image Benchmark

The social image benchmark assumes $\alpha = 1$. I consider two cases (1) $\beta \geq 1$ and (2) $\beta < 1$.

Proof.
Case 1. Assume $\beta \geq 1$. There is one Nash Equilibrium in pure strategies (T, F) because T is weakly dominant. The receiver's posterior belief that the message is correct trivially boils down to

$$\mu = \frac{1}{1+0} = 1.$$

Given belief μ and σ_R, the sender has no incentive to deviate from sending the correct message, since his expected payoff of sending an incorrect answer (1) is not larger than the expected payoff of sending the correct answer $(0 + \beta)$.

Case 2. Assume $\beta < 1$. The same analysis as in Case 1 of the previous proposition applies. Integrating the social image component β the best response correspondence of the sender is

$$\sigma_S^*(\mu) = \begin{cases} 0 & \text{if } \sigma_R < \frac{1+\beta}{n} \\ [0,1] & \text{if } \sigma_R = \frac{1+\beta}{n} \\ 1 & \text{if } \sigma_R > \frac{1+\beta}{n} \end{cases}$$

From inspection we see that the mixed strategy $(\sigma_S^*, \sigma_R^*) = (\frac{1}{n}, \frac{1+\beta}{n})$ with belief $\mu^* = \frac{1}{n}$ can be sustained as equilibrium strategy.

□

3.6.2 Pre-study

The pre-study was conducted after a completely unrelated individual decision-making experiment on time inconsistency in which subjects earned between 10 and 12.50 Euro. The pre-study consisted of three parts: (1) 50 questions, (2) belief elicitation about performance, (3) social status elicitation. English translations of all 50 questions can be found in Appendix 3.6.3. For each correctly answered question they received a prize of 6 cent, in total up to 3 Euro. In the belief elicitation, subjects were asked to indicate their belief about

their number of correctly answered questions. Subjects were paid 1 Euro in case their guess was equal or +/- 1 to the actual number of correct answers; and 1 Euros divided by the absolute estimation error if the estimate deviated by more than 1 answer. The social status elicitation consisted of two parts. On the first screen, subjects indicated how they assessed on average a person who is successful at answering the questions, on six different characteristics: intelligence, memory, success in studies and life, curiosity, openness for experiences and extraversion on a five-point Likert scale (image rating). On the second screen, subjects indicated the personal importance they attach to each characteristic (importance rating).

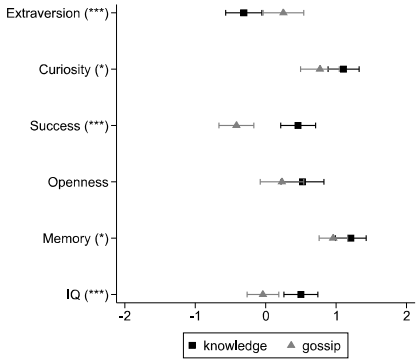


Figure 3.5: Social status rating of each characteristic in pre-study

Notes: The vertical lines depict 95% confidence intervals. Answers on the 5-point Likert scale corresponded to "low", "rather low", "neutral", "rather high", "high". Significance levels (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$) are indicated in parentheses. P-values of MW tests are as follows: IQ: $p = .003$, memory: $p = .072$, openness: $p = .243$, success: $p = .0$, curiosity: $p = .089$, extraversion: $p = .005$.

Figure 3.5 shows the image rating of the six characteristics in *high* and *low*. Figure 3.6 depicts the importance ratings of each characteristic. In terms of social image, there are statistically significant differences in four out of six categories with extraversion being the only characteristic that is more highly evaluated in *low*. In terms of importance, there are only two significantly different evaluations. Additionally, I construct an aggregate measure that controls for the individual importance of each characteristic. Denote the first measure by $norm_{ij}$ for each subject i and characteristic j and the second measure imp_{ij}. Taken these two measures, I calculate a social status score for each sub-

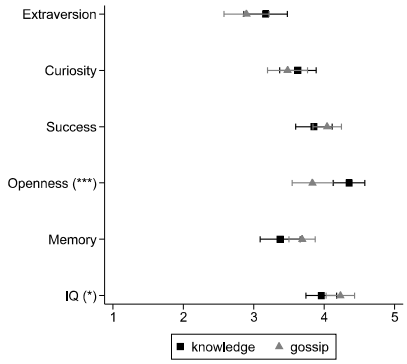


Figure 3.6: Importance rating of each characteristic in pre-study

Notes: The vertical lines depict 95% confidence intervals. Answers on the 5-point Likert scale corresponded to "unimportant", "rather unimportant", "does not matter", "rather important", "important". Significance levels (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$) are indicated in parentheses. P-values of MW tests are as follows: IQ: $p = .07$, memory: $p = .15$, openness: $p = .007$, success: $p = .405$, curiosity: $p = .619$, extraversion: $p = .187$.

ject $status_i = (\sum_{j=1}^{5} imp_{ij} * norm_{ij}) / \sum_{j=1}^{5} imp_{ij}$. It amounts to 0.61 in *high* and 0.27 in *low* (MW, $p = 0.0048$), which is a highly significant difference.

3.6.3 Multiple-Choice Questions

The tables 3.4 and 3.5 display all gossip and knowledge questions used in the pre-study and the corresponding correct answers. Tables 3.6 and 3.7 display which of the questions were additionally used in the sender-receiver game and in the post-experimental questionnaire, provide summary statistics for each question (share of answers chosen in percent), and the easiness level (which was shown to the subjects in the sender-receiver game).

Table 3.4: Questions in *low*

#	Question	Answer a)	Answer b)	Answer c)	Answer d)	Correct
1	Which actor plays Bilbo Baggins in The Hobbit?	Elijah Wood	Benedict Cumberbatch	Morgan Freeman	Martin Freeman	d
2	Which national-team footballer recorded the song Gute Freunde kann niemand trennen in 1966?	Franz Beckenbauer	Sepp Maier	Uli Hoeneß	Gerd Müller	b
3	Which brand is associated with the slogan "Freude am Fahren"?	Mercedes	BMW	Audi	Porsche	b
4	In which country was Irina Shayk, the ex-girlfriend of football star Cristiano Ronaldo, born?	Brasil	Portugal	Ukraine	Russia	d
5	In which country did Carlsberg beer originate?	Czech Republic	Denmark	Germany	Great Britain	b
6	Which of these countries provides no filming location for the series Game of Thrones?	Israel	Croatia	Malta	Northern Ireland	a
7	Which family has reigned in the Principality of Monaco for many decades?	Chevrier	Grimaldi	Pegues	Rozier	b
8	In which sporting discipline did Charlene Wittstock, Princess of Monaco, perform at international level?	Gymnastics	Swimming	Hockey	Diving	b
9	Which country is the drink Pernod from?	Italy	Portugal	France	Greece	c

10	Which postcode is associated with Beverly Hills in the eponymous series?	80210	80212	90211	90210	d
11	Who hosted the programme Nur die Liebe zählt until 2011?	Oliver Geissen	Stefan Raab	Jörg Pilawa	Kai Pflaume	d
12	What is Alan Harper's profession in Two and a Half Men?	Tax consultant	Chiropractor	Policeman	Accountant	b
13	Which German TV channel broadcasts the programme Shopping Queen?	RTL2	SuperRTL	Vox	Sat1	c
14	Which band has a red tongue as its logo?	Status Quo	AC/DC	The Doors	The Rolling Stones	d
15	Who left the boy band Take That in 1995?	Gary Barlow	Mark Owen	Jason Orange	Robbie Williams	d
16	Which city hosts the programme ZDF-Fernsehgarten?	Cologne	Düsseldorf	Mainz	Wiesbaden	c
17	In which city do police inspectors Frank Thiel und Karl-Friedrich Boerne work, in the German Tatort crime series?	Constance (Konstanz)	Munich	Münster	Cologne	c
18	Which drink is also referred to as "green fairy"?	Tequila	Waldgeist	Absinth	Kiwi schnapps	c
19	Which bank is associated with the slogan "Wir machen den Weg frei"?	Dresdner Bank	Deutsche Bank	Sparkasse	Volksbanken Raiffeisenbanken	d
20	What is the name of Lady Gaga's debut album?	The Frame	The Fake	The Fame	The Flame	c

21	What is the name of the winner of the first Big Brother series?	Jürgen	Percy	Alida	John	d
22	Which German soap opera boasts the most episodes broadcast?	Marienhof	Verbotene Liebe	GZSZ	Unter Uns	c
23	What is the name of Nina Hagen's daughter?	Cosma Shiva	Cosima Banghli	Nadia Rashnee	Ghislaine	a
24	What is the name of the Munich discotheque that belongs to the Käfer dynasty?	Hippodrome	P1	Omen	Tresor	b
25	Who plays Joey in the series Friends?	Matthew Perry	Tate Donovan	David Schwimmer	Matt Le Blanc	d
26	With whom did Jan Böhmermann host the radio programme Sanft & Sorgfältig?	Klaas Heufer-Umlauf	Joachim Winterscheidt	Olli Schulz	Charlotte Roche	c
27	Who has killed a White Walker in the series Game of Thrones?	Tyrion Lannister	Brienne of Tarth	Samwell Tarly	Ned Stark	c
28	What is the name of the beer brand the Simpsons drink?	Budweiser	Duff	Miller	Guinness	b
29	What is the name of the starfish who is Spongebob's best friend in the eponymous cartoon?	Johnny	Patrick	Sandy	Taddäus	b
30	Which country is Justin Bieber from?	Great Britain	Australia	Canada	USA	c
31	Which is James Bond's favourite cocktail?	Martini	Bloody Mary	Manhattan	Margarita	a
32	Which animal can one see whenever the online service Twitter crashes?	Bear	Whale	Fox	Lion	b

33	How often was the actress Elizabeth Taylor married?	Seven times	Six times	Five times	Eight times	d
34	Who won the contest Germany's Next Topmodel in 2015?	Vanessa Fuchs	Stefanie Giesinger	Anuthida Ploypetch	Kim Hnizdo	a
35	Who hosted the final series of Deutschland sucht den Superstar?	Michelle Hunziker	Nazan Eckes	Dieter Bohlen	Oliver Geissen	d
36	What kind of animal is Hein Blöd in Käpt'n Blaubär (Captain Bluebear)?	Horse	Rat	Dog	Bear	b
37	Which animal was the Italian cartoon character Calimero?	Brown bear	Black tomcat	Blue mouse	Black chick	d
38	Which of these actors has yet to win an Oscar?	Johnny Depp	Leonardo di Caprio	Christoph Waltz	Sean Penn	a
39	In the series Friends, Ross marries three times. With whom was he never married?	Monica	Carol	Emily	Rachel	a
40	How many children do Brad Pitt and Angelina Jolie have?	7	6	4	5	b
41	Tom Cruise is a member of which religion?	Scientology	Buddhism	Jehovah's Witnesses	Judaism	a
42	With which US State is the whiskey brand Jack Daniels usually associated?	Mississippi	Alabama	Tennessee	Georgia	c
43	Which alcoholic drink is used in the preparation of a Daiquiri cocktail?	Rum	Gin	Whiskey	Vodka	a
44	Which country won the 2016 Eurovision Song Contest?	Ukraine	Russia	Aserbaidschan	Netherlands	a

45	In the year of which animal are we currently, according to the Chinese horoscope?	Dog	Monkey	Pig	Horse	b
46	What is associated with the following slogan: "Da weiß man, was man hat"?	Persil	Ariel	Meister Proper	Perwoll	a
47	What was Lady Diana's profession before she married Prince Charles?	Nurse	Teacher	Nursery school teacher	Secretary	c
48	Where might one drink a "Weiße mit Schuss"?	Munich	Berlin	Hamburg	Düsseldorf	b
49	Who won the Oscar in the Best Actor category in 2016?	Matt Damon	Michael Fassbender	Matthew McConaughey	Leonardo di Caprio	d
50	What is the son of Prince William and his wife Catherine called?	William	Philip	George	Harry	c

Table 3.5: Questions in *high*

#	Question	Answer a)	Answer b)	Answer c)	Answer d)	Correct
1	What is the name of the mathematician credited with a famous concept in game theory, named after him?	Carl Friedrich Gauss	Alan Turing	Bernard Bolzano	John Nash	d
2	What does "intrinsic" mean?	Cunning	Of one's own accord	Turned inwards	Dreamy	b
3	What does Grand Marnier taste of?	Orange	Fig	Plum	Apricot	a
4	Where is the Taunus mountain range situated?	Hesse & Rhineland-Palatinate	Bavaria & Baden-Württemberg	Thuringia & Saxony	Lower Saxony	a
5	Julian Assange fled to an embassy of which country?	Venezuela	Ecuador	Sweden	Russia	b
6	In which epoch was Lessing's Nathan the Wise published?	Classicism	Romanticism	Realism	Enlightenment	d
7	Which river flows into the Rhine in Mannheim?	Jagst	Moselle	Isar	Neckar	d
8	What is the highest mountain in the European Union?	Mont Blanc	Zugspitze	Matterhorn	Etna	a
9	Who elects the German Chancellor (Bundeskanzler)?	Federal Council (Bundesrat)	Bundestag	Federal Assembly, or Bundesversammlung	Federal Government	b
10	What is a dividend?	The price of a share	A share proportion one owns	Earnings per share	Profit distribution per share	d
11	How many degrees does the sum of all angles have in a triangle?	380°	360°	180°	90°	c
12	What is cardamom?	A city in Armenia	A spice	Llama wool	A hormone	b
13	Which of these mobile internet networks has the fastest potential transmission rate?	3G	LTE	GPRS	EDGE	b
14	Which of the following is a term from chaos theory?	Butterfly effect	Eagle effect	Seagull effect	Bumblebee effect	a
15	Which gemstone is green?	Opal	Ruby	Emerald	Sapphire	c
16	What is the largest island on earth?	Greenland	Neu Guinea	Madagascar	Sumatra	a
17	What is the correct spelling of the German word for "apparatus"?	Aparatur	Aperatur	Apparatur	Apperatur	c
18	In which country is a straw hat worn in such a way that it is possible to tell the wearer's life situation?	Panama	Venezuela	Tunisia	Marocco	a
19	What is the capital of Turkey?	Ankara	Istanbul	Izmir	Antalya	a
20	Which SIM card format does not exist?	Mini SIM	Small SIM	Micro SIM	Nano SIM	b
21	What is the name of the French blue cheese that is both blue and green and made from raw sheep's milk?	Bavaria Blu	Adelöst	Danablu	Roquefort	d

22	What is the name of the Greek doctor whose professional ethics still apply today?	Damocles	Hippocrates	Diogenes	Aristotle	b
23	Who discovered the sea route to India?	Christopher Columbus	Ferdinand Magellan	James Cook	Vasco da Gama	d
24	What does the Pearl Index calculate?	Purity of diamonds	Inequality of wealth	Reliability of contraceptive methods	Cleanliness of water	c
25	Which of these islands is not North Frisian?	Sylt	Amrum	Föhr	Norderney	d
26	Who is considered the founder of evolutionary theory?	Konrad Lorenz	Iwan Pawlow	Charles Darwin	Gregor Mendel	c
27	What is understood by the word "recession"?	Economic upturn	Economic high	Depression	Economic downturn	d
28	Who was President of France before François Hollande?	Laurent Fabius	Marine Le Pen	Nicolas Sarkozy	Jacques Chirac	c
29	What is known as "Trisomy 21"?	Haemophilia	Cystic fibrosis	Brittle bone disease	Down syndrome	d
30	What is the capital of Hesse?	Frankfurt	Düsseldorf	Darmstadt	Wiesbaden	d
31	Which element does not belong to the inert gases?	Neon	Helium	Ozone	Argon	c
32	Who was the first American President?	Lincoln	Washington	Roosevelt	Franklin	b
33	What is, roughly, the circumference of the Earth?	60,000 km	40,000 km	30,000 km	20,000 km	b
34	When was the German Reich founded?	1871	1866	1848	1933	a
35	Which of the following is not one of the Balearic Islands?	Mallorca	Tenerife	Menorca	Ibiza	b
36	Which country was ruled by Frederick the Great?	Austria	The German Reich	Russia	Prussia	d
37	What was not the name of one of Jesus' 12 apostles?	Thomas	John	Balthasar	Peter	c
38	Which German Chancellor was in office longest?	Gerhard Schröder	Helmut Schmidt	Konrad Adenauer	Helmut Kohl	d
39	In which country can Apulia be found?	Norway	France	Italy	Israel	c
40	Which is the smallest Bundesland in Germany?	Bavaria	Hamburg	Bremen	Saarland	c
41	What is a persiflage?	A French drink	A type of plant	A style of painting	Mockery	d
42	Who painted "The Scream"?	Vincent van Gogh	Edvard Munch	Leonardo da Vinci	Paul Gauguin	b
43	Which European country does not have the Euro?	France	Portugal	Sweden	Slovakia	c
44	Whom did George W. Bush defeat in the 2000 Presidential election?	John Kerry	Marco Rubio	Bill Clinton	Al Gore	d
45	Who served as Foreign Minister under Gerhard Schröder?	Jürgen Trittin	Joschka Fischer	Thomas Oppermann	Sigmar Gabriel	b
46	What is the Roman sign for 50?	L	M	X	V	a
47	Which of the following countries is not a founding member of the European Union?	Italy	Luxemburg	Spain	Netherlands	c
48	What was the name of the Austrian heir to the throne, whose murder triggered the First World War?	Wilhelm II.	Franz Ferdinand	Charles V, Duke of Lorraine	Otto von Bismarck	b

49	Who wrote the dystopian tale Brave New World?	Karl Marx	George Orwell	H. G. Wells	Aldous Huxley	d
50	Who wrote the novel Perfume?	Patrick Süskind	Franz Kafka	Thomas Mann	Hermann Hesse	a

Table 3.6: Summary statistics of questions in *low*

#	Question	Use	a) in %	b) in %	c) in %	d) in %	Easiness in %
1	hobbit	T	37.50	25.00	0.00	37.50	40
2	lied	T	18.75	29.17	12.50	39.58	40
3	freude	T	4.17	43.75	37.50	14.58	45
4	ronaldo	T	14.58	8.33	33.33	43.75	45
5	carlsberg	T	31.25	47.92	12.50	8.33	50
6	thrones	T	54.17	14.58	16.67	14.58	55
7	grimaldi	T	16.67	58.33	20.83	4.17	60
8	charlene	T	20.83	58.33	20.83	0.00	60
9	pernod	T	10.42	18.75	68.75	2.08	70
10	beverly	T	8.33	8.33	10.42	72.92	75
11	liebe	T	14.58	0.00	6.25	79.17	80
12	twohalf	T	6.25	81.25	0.00	12.50	80
13	shopping	T	8.33	4.17	83.33	4.17	85
14	band	T	4.17	8.33	2.08	85.42	85
15	boyband	T	4.17	4.17	2.08	89.58	90
16	zdf	P	12.50	14.58	60.42	12.50	
17	tatort	P	4.17	12.50	70.83	12.50	
18	fee	P	0.00	68.75	27.08	4.17	
19	weg	P	6.25	12.50	4.17	77.08	
20	gaga	P	6.25	2.08	91.67	0.00	
21	brother	P	54.17	10.42	20.83	14.58	
22	seifenoper	P	20.83	16.67	56.25	6.25	
23	hagen	P	85.42	4.17	6.25	4.17	
24	kaefer	P	6.25	70.83	10.42	12.50	
25	joey	P	25.00	22.92	20.83	31.25	
26	boehmermann	P	4.17	14.58	60.42	20.83	
27	walker	P	10.42	12.50	52.08	25.00	
28	simpsons	P	2.08	93.75	2.08	2.08	
29	spongebob	P	0.00	100.00	0.00	0.00	
30	bieber	P	2.08	0.00	85.42	12.50	
31	bond		95.83	2.08	0.00	2.08	
32	twitterq		22.92	45.83	25.00	6.25	
33	taylor		6.25	47.92	43.75	2.08	
34	gntm		35.42	52.08	6.25	6.25	
35	superstar		16.67	39.58	22.92	20.83	

36	blaubaer	6.25	33.33	25.00	35.42
37	calimero	8.33	43.75	29.17	18.75
38	oscar	39.58	6.25	12.50	41.67
39	ross	33.33	16.67	22.92	27.08
40	kinder	25.00	27.08	16.67	31.25
41	religion	91.67	4.17	2.08	2.08
42	whiskey	2.08	10.42	85.42	2.08
43	daiquiri	27.08	35.42	4.17	33.33
44	eurovision	85.42	4.17	4.17	6.25
45	horoskop	18.75	41.67	25.00	14.58
46	waschmittel	33.33	29.17	18.75	18.75
47	diana	35.42	31.25	25.00	8.33
48	weisse	29.17	45.83	18.75	6.25
49	schauspieler	4.17	0.00	2.08	93.75
50	prinz	2.08	14.58	77.08	6.25

Notes: Abbrevations for use of questions are as follows: "T" means that question was used in the sender-receiver game in the treatment, "P" means that it was used as a post-experimental question. a), b), c) and d) in % indicates the percentage of answers for the corresponding answer item.

Table 3.7: Summary statistics of questions in *high*

#	Question	Use	a) in %	b) in %	c) in %	d) in %	Easiness in %
1	mathematiker	T	33.33	18.75	6.25	41.67	40
2	intrinsisch	T	12.50	41.67	39.58	6.25	40
3	marnier	T	43.75	16.67	14.58	25.00	45
4	taunus	T	43.75	14.58	33.33	8.33	45
5	assange	T	10.42	47.92	10.42	31.25	50
6	epoche	T	20.83	8.33	16.67	54.17	55
7	fluss	T	2.08	20.83	16.67	60.42	60
8	berg	T	58.33	25.00	14.58	2.08	60
9	bundeskanzler	T	22.92	68.75	4.17	4.17	70
10	dividende	T	0.00	6.25	20.83	72.92	75
11	winkelsumme	T	0.00	16.67	79.17	4.17	80
12	kardamon	T	4.17	79.17	10.42	6.25	80
13	internet	T	8.33	83.33	8.33	0.00	85
14	chaos	T	85.42	2.08	4.17	8.33	85
15	edelstein	T	6.25	0.00	87.50	6.25	90

16	insel	P	93.75	0.00	6.25	0.00
17	apparatur	P	22.92	4.17	54.17	18.75
18	strohhut	P	47.92	31.25	8.33	12.50
19	tuerkei	P	68.75	29.17	0.00	2.08
20	sim	P	8.33	85.42	2.08	4.17
21	kaese	P	29.17	8.33	4.17	58.33
22	eid	P	0.00	75.00	10.42	14.58
23	seeweg	P	18.75	27.08	27.08	27.08
24	pearl	P	25.00	37.50	14.58	22.92
25	nordfriesland	P	29.17	29.17	10.42	31.25
26	evolution	P	0.00	0.00	97.92	2.08
27	rezession	P	10.42	2.08	16.67	70.83
28	hollande	P	0.00	2.08	85.42	12.50
29	trisomie	P	8.33	0.00	4.17	87.50
30	hessen	P	27.08	2.08	10.42	60.42
31	edelgas		12.50	6.25	64.58	16.67
32	praesident		27.08	52.08	8.33	12.50
33	erdumpfang		37.50	52.08	10.42	0.00
34	dtreich		50.00	4.17	25.00	20.83
35	balearen		8.33	62.50	4.17	25.00
36	friedrich		8.33	16.67	4.17	70.83
37	apostel		16.67	2.08	79.17	2.08
38	amt		10.42	16.67	20.83	52.08
39	apulien		14.58	6.25	66.67	12.50
40	bundesland		2.08	20.83	50.00	27.08
41	persiflage		4.17	4.17	27.08	64.58
42	schrei		35.42	52.08	4.17	8.33
43	land		0.00	4.17	81.25	14.58
44	bush		20.83	0.00	27.08	52.08
45	schroeder		8.33	64.58	14.58	12.50
46	zeichen		54.17	39.58	2.08	4.17
47	eu		6.25	31.25	54.17	8.33
48	thronfolger		10.42	62.50	8.33	18.75
49	author		14.58	29.17	22.92	33.33
50	parfuem		62.50	12.50	14.58	10.42

Notes: Abbrevations for use of questions are as follows: "T" means that question was used in the sender-receiver game in the treatment, "P" means that it was used as a post-experimental question. a), b), c) and d) in % indicates the percentage of answers for the corresponding answer item.

3.6.4 Additional Tables

Table 3.8: Truth-telling and trust rates in *signaling*

	Truth-telling				Trust			
	High		Low		High		Low	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Actual rate	46.11	(0.23)	31.59	(0.21)	38.61	(0.33)	48.49	(0.33)
Senders' belief	41.39	(0.18)	33.33	(0.21)	45.28	(0.23)	37.97	(0.25)
Receiver's belief	31.94	(0.24)	39.13	(0.22)	40.28	(0.26)	40.29	(0.22)

Table 3.9: Truth-telling and trust rates in *no-signaling*

	Truth-telling				Trust			
	High		Low		High		Low	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Actual rate	30.91	(0.17)	33.17	(0.21)	51.3	(0.34)	55.56	(0.32)
Senders' belief	33.86	(0.12)	34.53	(0.12)	34.78	(0.2)	38.33	(0.23)
Receiver's belief	35.06	(0.13)	38.14	(0.12)	48.11	(0.27)	45.83	(0.28)

3.6.5 Instructions

This appendix presents paper instructions for the sender-receiver game used in the experiment 3.6.5, the instructions for the social status elicitation in the pre-study 3.6.5 and in the experiment 3.6.5.

Instructions for Sender-Receiver Game with Endogenous Information

In the following a translation of the original German instructions is shown, first the instructions for the *signaling-low* treatment (denoted by TG in the left upper corner), then for the *no-signaling-low* treatment (denoted by UG).

The instructions for the *high* treatments only differ with regards to the question that was used in the example. The question in the *high* treatment reads as follows:
Who wrote the dystopian tale "Brave New World"?

(a) Karl Marx

(b) George Orwell

(c) H. G. Wells

(d) Aldous Huxley

TG

General Instructions for Participants

You are about to take part in a study. If you read the following instructions carefully, then you can earn money – depending on your decisions and on those of the other participants. It is therefore crucial that you read these instructions carefully. You will remain anonymous for the duration of the entire experiment. You will neither know the identity of the other participants, nor will they know yours.

Communication is absolutely forbidden during the experiment. If you have any questions, please ask us directly. Disobeying this rule will lead to immediate suspension from the experiment and all payments.

In a previous study at the BonnEconLab, participants were asked questions on a variety of topics, and four answers (a, b, c, or d) were possible. These participants were able to earn money for each correctly answered question.

Procedure

The experiment consists of 15 rounds. At the beginning of the experiment, all participants are randomly divided into two participant types, participant A and participant B. Your participant type (A or B) will remain the same for all 15 rounds.

At the beginning of each round, groups of two players are randomly formed, each group consisting of a participant A and a participant B. The constellation of each group of two is drawn anew for every round. Neither you nor the other participant in your group of two will receive information about the other player. You are completely anonymous.

In each round, both participants (A and B) are shown a question from the previous study. Both are told how many percent of participants from the previous study answered the same question correctly (rounded to 5%). In addition, participant A is shown the possible answers (i.e., which concrete answers are behind the possible entries a, b, c, and d).

Participant A then sends a message to participant B. Participant A may choose between a, b, c, or d for this. Participant B receives participant A's message and then decides whether to choose a, b, c, or d.

In the experiment, the participants A first make their decisions for all 15 rounds; only once this has happened do all participants B make their decisions for all 15 rounds.

Payment

If participant B's decision actually **corresponds** to the correct answer to the question, then the payoff is as follows:

- Participant A: **6 Euro**,
- Participant B: **9 Euro**.

If participant B's decision **does not correspond** to the correct answer to the question, then the payoff is as follows:

- Participant A: **9 Euro**,
- Participant B: **6 Euro**.

TG

Example

The following question was in the previous study: Who won the competition "Germany's Next Topmodel" in 2015?

a) Vanessa Fuchs
b) Stefanie Giesinger
c) Anuthida Ploypetch
d) Kim Hnizdo

In today's experiment you would receive the following information.

Screen **Participant A**	Screen **Participant B**
Question 1: Who won the competition "Germany's Next Topmodel" in 2015? a) Vanessa Fuchs b) Stefanie Giesinger c) Anuthida Ploypetch d) Kim Hnizdo The question was answered correctly in the previous study by (in %): **35** Your message to participant B: ○ a ○ b ○ c ○ d	Question 1: Who won the competition "Germany's Next Topmodel" in 2015? The question was answered correctly in the previous study by (in %): **35** Participant A's message: **<a, b, c, or d>** Your decision: ○ a ○ b ○ c ○ d

Results

At the end of the experiment, you will receive the results of the 15 rounds. You will be told:

- the correct answer to the question: a, b, c, or d;
- the message sent by participant A;
- participant B's decision;
- the payments for participant A and participant B.

Payment

At the end of the experiment, one of the 15 rounds is chosen randomly as payoff-relevant for each participant.

UG

General Instructions for Participants

You are about to take part in a study. If you read the following instructions carefully, then you can earn money – depending on your decisions and on those of the other participants. It is therefore crucial that you read these instructions carefully. You will remain anonymous for the duration of the entire experiment. You will neither know the identity of the other participants, nor will they know yours.

Communication is absolutely forbidden during the experiment. If you have any questions, please ask us directly. Disobeying this rule will lead to immediate suspension from the experiment and all payments.

In a previous study at the BonnEconLab, participants were asked questions on a variety of topics, and four answers (a, b, c, or d) were possible. These participants were able to earn money for each correctly answered question.

Procedure

The experiment consists of 15 rounds. At the beginning of the experiment, all participants are randomly divided into two participant types, participant A and participant B. Your participant type (A or B) will remain the same for all 15 rounds.

At the beginning of each round, groups of two players are randomly formed, each group consisting of a participant A and a participant B. The constellation of each group of two is drawn anew for every round. Neither you nor the other participant in your group of two will receive information about the other player. You are completely anonymous.

In each round, both participants (A and B) are shown a question from the previous study. Both are told how many percent of participants from the previous study answered the same question correctly (rounded to 5%). *In addition, participant A is given information about what a randomly drawn player chose in the previous study (a, b, c, or d).* However, neither of the two players sees the possible answers (i.e., which concrete answers are behind the possible entries a, b, c, and d).

Participant A then sends a message to participant B. Participant A may choose between a, b, c, or d for this. Participant B receives participant A's message and then decides whether to choose a, b, c, or d.

In the experiment, the participants A first make their decisions for all 15 rounds; only once this has happened do all participants B make their decisions for all 15 rounds.

Payment

If participant B's decision actually **corresponds** to the correct answer to the question, then the payoff is as follows:

- Participant A: **6 Euro**,
- Participant B: **9 Euro**.

If participant B's decision **does not correspond** to the correct answer to the question, then the payoff is as follows:

- Participant A: **9 Euro**,
- Participant B: **6 Euro**.

UG

Example

The following question was in the previous study: Who won the competition "Germany's Next Topmodel" in 2015?

a) Vanessa Fuchs
b) Stefanie Giesinger
c) Anuthida Ploypetch
d) Kim Hnizdo

In today's experiment you would receive the following information.

Screen **Participant A**	Screen **Participant B**
Question 1: Who won the competition "Germany's Next Topmodel" in 2015? The randomly selected participant from the previous study chose**: <a, b, c, or d>** The question was answered correctly in the previous study by (in %): **35** This means the answer from the previous study is correct with the following probability (in %): **35** Your message to participant B: ○ a ○ b ○ c ○ d	Question 1: Who won the competition "Germany's Next Topmodel" in 2015? The question was answered correctly in the previous study by (in %): **35** Participant A's message: **<a, b, c, or d>** Your decision: ○ a ○ b ○ c ○ d

Results

At the end of the experiment, you will receive the results of the 15 rounds. You will be told:

- the correct answer to the question: a, b, c, or d;
- the information received by participant A;
- the message sent by participant A;
- participant B's decision;
- the payments for participant A and participant B.

Payment

At the end of the experiment, one of the 15 rounds is chosen randomly as payoff-relevant for each participant.

Instructions for Social Status Elicitation in Pre-Study

These questions were presented on the participant's screen.

Screen 1
Please assess the significance of correct answers in the prior questions.
The table below depicts a list of characteristics.
Please indicate for each characteristic how you evaluate each one on average for people who score well on the prior questions. Choose between "low", "rather low", "neutral", "rather high", "high". Please choose as option the one you think is chosen by most participants.
... Intelligence quotient
... Memory
... Success in studies and life
... Curiosity
... Openness for new experiences
... Extroverted personality

Screen 2
How important do you find it that other people perceive you as someone with the following characteristics?
(unimportant – rather unimportant – does not matter – rather important – important)
... Intelligence quotient
... Memory
... Success in studies and life
... Curiosity
... Openness for new experiences
... Extroverted personality

Instructions for Social Status Elicitation in Experiment

To improve the clarity of the questions, some language changes were implemented in the experiment. The *importance* question from the pre-study was split up into two questions (perception of each characteristic and relative importance). One characteristic, conviviality, was added in the experiment. The

first assessment question was incentivized as in Krupka and Weber (2013) and gave a prize of 2 Euro when the subject's answer matched the modal answer. One of the seven characteristics was randomly chosen for payment. In the pre-study subjects were told to indicate the answer that would be chosen by most subject. These questions were presented on the participant's screen.

Screen 1
The following table shows a list of characteristics. Please estimate for each of the characteristics the relationship between the number of correctly solved questions and the respective characteristic.

Choose whether you perceive the relationship as "negative", "slightly negative", "no connection", "weakly positive" or "positive". To indicate your answer, click on the appropriate box.

Example: Relationship between number of questions solved and health

1. A positive relationship exists when those who solved many questions correctly, are on average more healthy.
2. No relationship exists when those who solved many questions correctly, are on average not particularly healthy nor particularly unhealthy.
3. A negative correlation exists when those who solved many questions correctly, are on average unhealthy.

At the end of the experiment, one of the characteristics will be randomly selected. For this characteristics we determine the option that was selected by most participants.

If you have chosen the same answer as most other participants, you will receive in addition to the other payments 2 Euro.

Relationship between number of questions solved and the following characteristics
... Intelligence quotient

... Memory
... Success in studies and life
... Curiosity
... Openness for new experiences
... Extroverted personality
... Conviviality

Screen 2
How do you feel it when you are perceived by others as someone with the following characteristics? (negative – rather negative – neutral – rather positive – positive)
... Intelligence quotient
... Memory
... Success in studies and life
... Curiosity
... Openness for new experiences
... Extroverted personality
... Conviviality

Screen 3
How important are these characteristics to you? You have 100 points. Please distribute these 100 points on the characteristics. The more important the particular characteristic is to you, the more points you should distribute to this characteristic.
... Intelligence quotient
... Memory
... Success in studies and life
... Curiosity
... Openness for new experiences
... Extroverted personality
... Conviviality

Chapter 4

Institutional Endogeneity and Third-Party Punishment in Social Dilemmas

4.1 Introduction

Institutions are a crucial determinant of social interaction outcomes. In situations in which self-interest clashes with collective interest - so-called social dilemmas - human societies have developed a number of institutional arrangements that mitigate the inherent free-rider problem. The third-party enforcement of social norms is one such arrangement that has shown to be successful in enhancing cooperation (Charness, Cobo-Reyes, and Jiménez, 2008; Lergetporer et al., 2014). In this paper we study the extent to which the process that generates a third-party punishment institution influences its severity and how the affected individuals respond to it. In particular, we compare settings in which groups select third-party sanctions through majority voting with settings in which sanctions are exogenously put in place. This allows us to investigate whether third-parties who have been introduced through majority voting enforce cooperation norms differently than those who have been exogenously appointed.

The available empirical evidence shows that endogenous punishment institutions lead to more cooperative behavior than their exogenous counterparts. This phenomenon is typically referred to as the 'endogeneity effect'.

In particular, institutions tend to be more effective in increasing cooperation when individuals can determine their implementation through majority voting (e.g., Tyran and Feld, 2006). In principle, this effect could be due to the self-selection of cooperative individuals into their preferred institution, to the cooperation signal inherent to the voting outcome, or to the concession of participation rights to individuals. Several experimental papers show that endogeneity increases cooperation even after controlling for the selection and signaling channels (e.g., Dal Bó, Foster, and Putterman, 2010; Kamei, Putterman, and Tyran, 2015). This regularity is referred to as the 'endogeneity premium': allowing groups to adopt sanction or reward schemes drives an increase in cooperative behavior.

While the existing studies on centralized institutions involve punishment levels that are predetermined and executed by an automatic mechanism, in reality punishment is the product of human judgment and is frequently administered by third parties. In most cases punishment is carried out under more or less established sets of rules (social norms, legal rules, etc.), but leeway is granted to the authority by whom it is administered. One prominent example is the judicial system: judges are bound by law but decide cases at their own discretion to some extent. Authorities also have plenty of scope to decide about the extent to which they punish non-cooperative behavior in less formalized settings (in the workplace, in the classroom or in any kind of self-organizing community). Our framework captures both dimensions of the typical punishment institution: the authority that applies punishment is free to decide on its extent within a set of rules that limit the severity of sanctions.

Several real-life instances exist in which different institutional procedures are used to select the officials in charge of administering justice and enforcing the law. Lay jurors are a case in point, as they are randomly selected in some countries (with a Common Law tradition), while in other countries they are appointed (most continental European countries), and yet in others they are directly elected by citizens (some cantons in Switzerland).[1] In the United States, judges at all levels of the judiciary are appointed in some states but elected in others, and this seems to influence their judicial decisions (e.g., Hanssen, 1999). The same is true for public prosecutors, with Rasmusen, Raghav, and Ramseyer (2009) suggesting that elections cause prosecutors to aim at higher conviction rates. At the law enforcement level, sheriffs and chiefs of police often share an overlapping mandate. Whereas the overwhelming majority of

[1]For an overview of juror selection methods see Jackson and Kovalev (2006).

sheriffs are elected by their constituencies, all police chiefs are appointed. In general there is very little empirical evidence on how the sanctioning behavior (e.g., conviction rates, severity of penalty) of lay jurors, judges, prosecutors or law enforcement officials depends on their selection procedure. Besides the lack of data availability and severe restrictions on data use (for data issues of jury trial data see e.g., Anwar, Bayer, and Hjalmarsson, 2015), a major problem for the empirical analysis is the endogeneity of the selection procedure, as different groups tend to adopt different institutional arrangements. The causal effect of endogenous institutional choice cannot be disentangled from the characteristics of the individuals who make the choice and the profile of the official who administers the sanctions.

Using a laboratory experiment, we investigate whether institutional endogeneity per se matters for the severity of third-party punishment in a social dilemma. After gaining experience in a multi-person prisoner's dilemma, our subjects vote whether they wish to play the same version of the game or a modification that allows for a third-party to punish defectors. The punishment decisions of an elected third-party punisher are compared to those of a third-party punisher who has been randomly appointed. We provide a theoretical framework based on the outcome-based social preference model of Charness and Rabin, 2002 to explain the punishment decisions of the differently selected third-party punishers. Our experimental design allows us to control for the mentioned selection and signaling channels inherent to institutional choice settings.

For groups where the majority favors a punishment institution, we find that punishment amounts to an average of 40.2% of the maximum punishment level in the exogenous institutional setting and 14.4% in the endogenous one. The difference in punishment severity may be explained by differences in its expected effectiveness. Indeed, assigned punishment points are significantly more successful in getting defectors to contribute in the endogenous institutional setting. That is, we show that endogenous third-party sanctions are less harsh and more effective than exogenous ones, all else equal. While endogenous institutions start out generating higher public good contributions, confirming the existence of an endogeneity premium, over time the more severe punishment implemented in the exogenous case increases contributions beyond those of the endogenous counterpart. Overall efficiency is not different across endogenous and exogenous institutions, yet the required punishment levels are significantly lower in the endogenous setting.

Our results offer an important insight for institutional choice. Punishment by a third-party is less severe when the sanctioning institution is adopted democratically, but punishment is also more effective. That is, endogenously selected sanctions are more persuasive in changing behavior than exogenously imposed ones. We further contribute to the literature by showing that voting over sanctions does not only affect the behavior of the parties who take part in the procedure, but also the decisions of the individuals who are responsible for administering them.

4.1.1 Related Literature

Many studies have analyzed the effectiveness of punishment institutions in enhancing and sustaining cooperation in social dilemmas (e.g., Andreoni, Harbaugh, and Vesterlund, 2003; Fehr and Gächter, 2000; Ostrom, Walker, and Gardner, 1992). A burgeoning literature has explored which conditions are most conducive to cooperation, e.g., the cost-to-effectiveness ratio of punishment, group size, and whether punishment or reward systems perform better (for an overview see Chaudhuri, 2011).

Recently, several authors have investigated the effectiveness of endogenous punishment institutions, focusing on two types of punishment regimes: *centralized formal* and *decentralized informal*. Centralized formal sanction mechanisms automatically reduce the payoff of defecting players by a certain amount. The literature has studied both costless (e.g., Tyran and Feld, 2006) and costly regimes (e.g., Markussen, Putterman, and Tyran, 2014). In costly regimes participants pay a fixed cost to have the scheme in place. Decentralized informal peer-to-peer punishment provides group members with the option to punish each other at a cost. Both the punishing and the punished subjects pay the cost, and typically the cost paid by the punisher is lower. Endogenously implemented centralized formal punishment regimes (Dal Bó, Foster, and Putterman, 2010; Kamei, Putterman, and Tyran, 2015; Markussen, Putterman, and Tyran, 2014; Tyran and Feld, 2006) and decentralized informal peer-to-peer punishment regimes (Markussen, Putterman, and Tyran, 2014; Sutter, Haigner, and Kocher, 2010) have both proven to be more effective than their exogenous counterparts.

Tyran and Feld (2006) were the first to report on the existence of the so-called endogeneity effect, showing that cooperation in a public good game is

higher when the punishment institution is enacted through a majority voting procedure rather than by the experimenter.[2] Sutter, Haigner, and Kocher (2010) confirm this regularity. In their experiment, subjects can choose whether to add a peer-to-peer sanction or reward scheme to a standard voluntary contribution mechanism (VCM). Under both schemes cooperation is found to be higher when the implementation is endogenous. In Markussen, Putterman, and Tyran (2014) subjects choose between costly formal sanctions, peer-to-peer sanctions and no sanctions. With experience subjects come to prefer peer-to-peer punishment, which they manage to implement efficiently. Both sanctioning institutions are more efficient when chosen collectively by majority vote than when exogenously implemented.

The effect of selecting the punishment institution through majority voting on cooperation may result from either the endogenous process itself or from side effects that the endogenous process brings about, namely self-selection and signaling. For instance, self-selection of cooperative individuals into the same institution could account for the observed higher cooperation. Groups that implement punishment may consist of participants whose preferences differ from those that choose not to implement it. In addition, the vote for the punishment institution can serve as a signaling or coordination device. That is, by voting for a certain institution participants signal their willingness to cooperate. This induces participants to infer each others' intentions from the voting outcome and to cooperate more. While some of the previous studies discuss and partially address the signaling and self-selection issues, the seminal mechanism proposed by Dal Bó, Foster, and Putterman (2010) manages to isolate the pure impact of endogeneity on cooperation. In their experiment, groups can vote on whether to interact in an environment with or without sanctions. The mechanism consists of a random draw that may overrule the group vote, followed by another random draw that implements one of the two environments in case the vote outcome was overruled. Controlling for self-selection through the comparison of groups that vote identically but differ on whether the choice was endogenously or exogenously implemented, the authors find a significant difference in cooperation rates. This constitutes evidence of

[2]Prior to Tyran and Feld (2006), endogenous choice of institutions in collective action scenarios was studied through mechanisms other than voting. For instance, Yamagishi (1986) investigate the endogenous funding of an exogenously available punishment mechanism, Ostrom, Walker, and Gardner (1992) analyze the combined effects of communication and voting, Gürerk, Irlenbusch, and Rockenbach (2006) and Nicklisch, Grechenig, and Thöni (2016) allow subjects to endogenously sort into different institutions by voting with one's feet.

an endogeneity premium.[3]

In a closely related line of research, several papers study which features of punishment institutions are likely to affect their perceived legitimacy, e.g., the compensation of punishers (Dickson, Gordon, and Huber, 2015), the accuracy of information available to punishers (Dickson, Gordon, and Huber, 2009) and the selection procedure for punishers (Baldassarri and Grossman, 2011; Grossman and Baldassarri, 2012). The latter two studies are closest to this paper. In a lab-in-the field experiment, the authors compare the punishment and contribution patterns under elected and randomly appointed third-parties. They show that groups with an elected punisher contribute on average more than those with a randomly selected one, and that elected punishers tend to sanction higher contributions than appointed ones. As contributions differ across treatments, the comparison of punishment choices between treatments is however limited. Furthermore, in the election treatment subjects can elect a punisher based on his visible characteristics, which gives rise to selection effects. Subjects are found to predominantly elect wealthy, highly educated males.[4]

Our experimental study makes a novel contribution to the literature by investigating whether the implementation procedure of a third-party punishment institution per se affects the severity of punishment and the resulting cooperation patterns. The extant evidence suggests that the impact of *endogeneity* on cooperation is a behavioral regularity in several settings, while the impact on punishment is yet unclear. To answer our research question we employ an experimental setting with centralized punishment whose severity is chosen by a third-party and then repeatedly applied. While a few existing studies allow for the endogenous choice of punishment levels, they do so in a decentralized

[3]Kamei, 2016 and Chen, 2014 use the same mechanism in their experimental design and replicate this regularity. In Kamei (2016) subjects play two public good games simultaneously. A non-deterrent centralized sanction scheme can be endogenously implemented in one game, whereas a random draw exogenously implements it in the other game. He finds significant evidence of an endogeneity premium in the endogenous game and positive spillover effects to cooperation in the exogenous game. Chen (2014) investigates the endogeneity premium in an experiment where subjects vote on non-deterrent formal sanctions in the absence and presence of peer-to-peer sanctions.

[4]The authors show that the recorded socio-demographic characteristics do no correlate with players' contributions and do not affect the response to punishment. Nonetheless subjects may respond to some visible characteristics of the punisher that are not collected as data. The authors further find that there are no differences in punishment behavior when they control for the divergence in contribution distributions.

peer-to-peer punishment setting (Markussen, Putterman, and Tyran, 2014; Sutter, Haigner, and Kocher, 2010). With several possible (peer) punishers it is not possible to identify the impact of endogeneity on punishment, as individuals' beliefs about the others' punishment behavior are crucial for the own punishment decision. Furthermore, since in the above-mentioned studies punishment decisions are taken in every round of the game, contribution choices and punishment choices can simultaneously affect each other.

To answer our research question, it is important to study third-party punishment instead of peer-to-peer punishment as the motives of unaffected outsiders reflect most closely the ones of authorities (e.g., jurors, judges and sheriffs). Unlike Baldassarri and Grossman, 2011 and Grossman and Baldassarri, 2012, who focus on the selection method of a third-party punisher, we are interested in how the institutional legitimacy of third-party punishment is influenced by the implementation process, and as such abstract from the role of personal characteristics of the punisher. Our experimental design further allows us to isolate the pure effect of endogeneity on punishment as it controls for selection and signaling effects.

The remainder of this paper is structured as follows. Section 4.2 summarizes the experimental design and procedures. In Section 4.3 we discuss predictions for punishment and contribution behavior. Section 4.4 includes the results and Section 4.5 concludes.

4.2 Design

At the beginning of the experiment subjects are randomly divided into groups of four. Two different roles are assigned within a group. Three group members are A-type subjects and one is a B-type subject. The experiment consists of two parts. A-types interact in a social dilemma in both parts. After part 1, A-types decide through majority voting in what institutional setting they want to interact in part 2. Part 2 can either be identical to part 1 or modified to allow for third-party punishment, to be administered by a B-type. Subjects know that the experiment comprises two parts, but only receive instructions for the second part after the first one is completed. Types are fixed throughout the experiment, but subjects are re-matched after part 1. We employ a perfect strangers protocol such that no subject is part of the same group in parts 1 and 2. Subjects are informed beforehand that 1 of the 20 periods from each part

will be randomly picked for payment at the end of the experiment. Earnings in the experiment are expressed in points, which are converted to Euro at the rate of 0.05 Euro per point. The sequence of the experiment is depicted in Figure 4.1.

4.2.1 Part 1

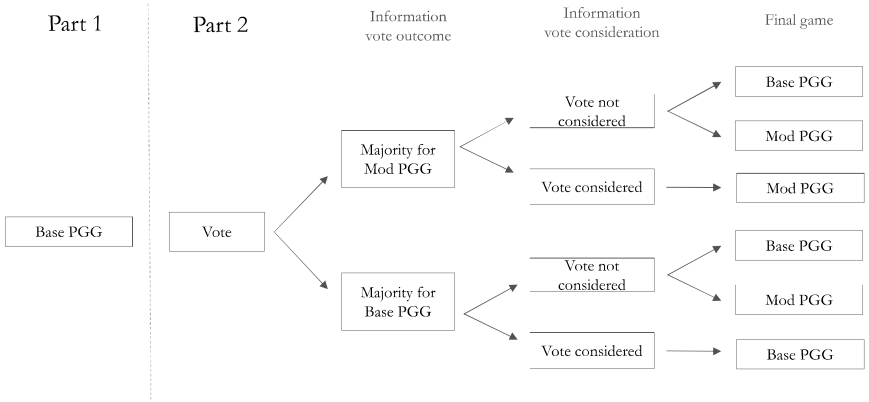
In the first part of the experiment the A-types play a 3-person prisoner's dilemma, which is equivalent to a public good game (PGG) with binary contribution choices. We stick to the latter terminology. The PGG is played for 20 periods with constant group composition. Part 1 is meant to familiarize subjects with the game and to allow them to gain experience such that they can make an informed voting decision.[5] Each A-type receives an endowment of $E_A = 70$ points in each period, which he or she can allocate to the group account ($c_i = 1$) or to the private account ($c_i = 0$). The A-types' income from the group account is the sum of all group members' contributions, $G = \sum_{i=1}^{3} c_i$, multiplied by $\alpha = 0.6$, the so-called marginal per capita return (MPCR). This results in the payoff function:

$$\pi_{A_i} = E_A(1 - c_i + \alpha G) \tag{4.1}$$

After each period, A-types learn the number of group members who contributed to the group account and their own period income. The B-types receive a fixed endowment of E_B=153 points (for reasons of consistency with part 2, which will be explained below). They are asked to indicate their beliefs about the A-types' behavior in the PGG. In particular, they state their expected distribution regarding the four possible outcomes in their group, i.e., in how many periods 0, 1, 2 or 3 A-types will contribute to the group account. The distribution entries must sum up to 20, the total number of periods (see Appendix 4.6.2 for details on the belief elicitation questions). For each correct entry the B-type receives 10 points, which means that in part 1 B-types can earn up to 40 points on top of the 153 points.

[5]Prior research has shown that inexperienced participants prefer environments without punishment. After accumulating some experience, however, subjects' preferences reverse and punishment institutions gain support (Botelho et al., 2005; Ertan, Page, and Putterman, 2009; Gürerk, Irlenbusch, and Rockenbach, 2009; Markussen, Putterman, and Tyran, 2014).

Figure 4.1: Sequence of the experiment



Notes: Only A-types participate in the Base PGG in Part 1. The dashed line between Part 1 and 2 depicts the perfect stranger rematching of groups between both parts. In Part 2 all the members of a group (the three A-types as well as the B-type) receive information on the vote outcome, vote consideration and the final game.

4.2.2 Part 2

As in part 1, the three A-types interact repeatedly for 20 periods. They play either the public good game without punishment (base PGG), identical to the one in part 1, or a modified public good game with third-party punishment (modified PGG). We will first explain the modified PGG, and afterwards the voting procedure and the process that determines which of the two games is implemented.

The modified PGG

In the modified PGG a third-party punishment regime is in place. Each B-type receives an endowment of $E_B = 153$ points. He or she can assign up to a maximum of 9 deduction points to each A-type who does not contribute to the group account, henceforth referred to as a 'defector'. The 153 points equal the full cooperation payoff of A-types (126 points) plus the maximum number of deduction points that can be assigned (27 points). The B-type cannot discriminate between defectors, i.e., all defectors incur the same amount

of punishment. We apply the strategy method to the third-party punisher's decision and ask him or her to indicate the amount of deduction points per defector conditional on the number of defectors, which we denote by m. The vector $\mathbf{d} = (d_1, d_2, d_3)$ denotes the number of punishment points assigned by the B-type to the defectors in each of the three possible cases ($m = 1, 2, 3$). Deduction points must be an integer between 0 and 9. We follow the literature (e.g., Fehr and Fischbacher, 2004a and Fehr and Fischbacher, 2004b) in that each assigned deduction point leads to a threefold reduction of a defector's income. The resulting payoff function for the A-types is:

$$\pi_{A_i} = \begin{cases} \alpha G E_A, & \text{if } c_i = 1 \\ (1 + \alpha G) E_A - 3d_m, & \text{if } c_i = 0 \end{cases} \tag{4.2}$$

Assigned deduction points lead to a one-to-one reduction of the B-type's income. Punishment costs are thus small relative to the B-type's endowment. For $d_m = 0$ the function conforms to that in Equation 4.1 for the base PGG. The resulting payoff function for the B-type is:

$$\pi_B = E_B - m d_m. \tag{4.3}$$

Given that we want to study the impact of endogeneity on cooperation enforcement we exclude the possibility to punish cooperators. In our experiment the punishment institution should unambiguously serve as a tool to foster cooperation. The exclusion of 'anti-social' third-party punishment also makes the institution more attractive to cooperative A-types and is closer to real-world applications. Central authorities, e.g., judicial systems, can only punish those who violate rules. Limiting punishment to a maximum of 9 deduction points allows us to restrict it to being non-deterrent. That is, the maximum number of points that can be deducted from an A-type's period income (27=9*3) is smaller than the gain that defection brings about (28 points). That way we can make sure to have a social dilemma, which would not be the case if punishment was deterrent (i.e., higher than 9 points). In other words, with non-deterrent sanctions A-types have no material incentive to cooperate while with deterrent sanctions cooperation would become the best response. In addition, ceiling effects could arise in such case.[6]

[6]For example, Tyran and Feld, 2006 observe a 93% cooperation rate under an exogenous deterrent sanction regime already.

As mentioned, the punishment vector is elicited using the strategy method and it is applied throughout the entire 20 periods. In each period the actual number of defectors determines which punishment decision applies. This design has the clear advantage that we can isolate the effects of the punishment decision on cooperation behavior and can exclude any endogeneity issues that would arise if contribution and punishment decisions could simultaneously affect each other. In other words, certain treatments could lead to 'back-' or 'front-loading' of punishment, e.g., an endogenously selected third-party punisher could adopt a lenient stance in the beginning but becomes harsher later on.[7] Another advantage of applying the strategy method is that it induces subjects to make deliberate decisions (Brandts and Charness, 2011) that are unaffected by emotions that would result from observed contribution behavior.

We elicit the B-types' punishment decisions after they have received information on the vote outcome and the vote consideration, i.e., B-types know whether a majority voted in favor of the modified PGG and whether the modified PGG was introduced as a consequence of the majority vote outcome or of a random draw. While the B-types decide on the punishment vector, we elicit the corresponding punishment beliefs from the A-types (see Appendix 4.6.2). An A-type earns 10 points for a correct belief in each of the three punishment vector entries. The modified PGG then starts. At the beginning of the second period the punishment vector is revealed to the A-types in order to avoid that uncertainty is resolved differently across treatments, as some cooperation outcomes may be more likely to occur in certain treatments. This could influence behavior in the PGG. The A-types' first period contributions are thus unaffected by the actual punishment vector or other group members' contributions. This is a deliberate design choice that allows us to assess the influence of the institution selection procedure on initial cooperation. As in the base PGG, we elicit the B-type's beliefs about A-types' behavior while they play the PGG. Furthermore the B-type is asked to fill out a questionnaire on his or her choices (see Appendix 4.6.2). At the end of the 20 periods the B-type receives information about the A-types' contributions and the resulting payoffs in each period.

[7] Tan and Xiao (2014) study the difference between ex-ante and ex-post punishment decisions in a one-shot prisoner's dilemma for individuals and groups acting as third parties. Punishment decisions do not vary for individual punishers, while groups punish more ex ante than ex post. Allowing punishment to react to observed cooperation levels in a multi-period setting is certainly an interesting research question, which can be addressed by future work.

Voting and vote consideration

The A-types are asked to decide via majority voting whether the base PGG or the modified PGG should be implemented in part 2. After all subjects cast their vote, a random mechanism determines whether the group's vote outcome is considered. With probability $p_v = 0.5$ the votes are considered and the majority vote determines which game is implemented, leading to what we call the endogenous institutional setting (Endo). With probability $1 - p_v = 0.5$ the votes are not considered and the computer randomly decides which game is implemented, leading to what we call the exogenous institutional setting (Exo). In Exo the modified PGG is implemented with probability $p_r = 0.9$ and the base PGG is implemented with probability $1 - p_r = 0.1$. The actual probabilities are not revealed to subjects but they are aware of the procedure. All subjects, A-types as well as the B-type, learn what the majority of the A-types in their group voted for, whether the votes are considered, and which game will be implemented.[8]

The random vote overrule procedure is taken from Dal Bó, Foster, and Putterman, 2010 and makes it possible to exclude selection and signaling effects from the results. Without this procedure there would be an asymmetry between treatments, i.e., if only subjects in our Endo treatment were allowed to vote. First, a vote in favor of the modified PGG signals a preference for cooperation, which may in turn affect the B subjects' punishment behavior as well as the other A subjects' willingness to cooperate (*signaling effect*). Second, cooperative behavior after a positive vote may be attributed to a *selection effect* since groups would be composed of subjects with identical institutional preferences. In other words, those who vote for modified PGG may be more likely to contribute to the group account than those who voted for the base PGG (see Tyran and Feld, 2006 and Dal Bó, Foster, and Putterman, 2010 for a more detailed discussion). The fact that all subjects may vote, in combination with the vote overrule procedure, allows us to control for group composition effects and to keep information about A-types' preferences constant across treatments.

Within each treatment, punishment may be implemented or not. In the Endo treatments the final institutional arrangement is the one decided by the

[8]Groups are not informed about individual votes as this would stress the signaling content of the vote outcome, and require us to compare groups with the same vote outcome across treatments, therefore reducing the statistical power of our analysis.

majority, so that two possible conditions may occur. In the Exo treatments, however, the opposite of what the majority voted for may be implemented. Thus, four possible Exo treatment conditions may occur. Table 4.1 lists all treatment conditions and the corresponding number of observations in our experiment.

Table 4.1: Conditions per treatment and observation numbers

Vote considered	Majority Mod PGG	Punish-ment	Abbre-viation	A-Types	B-Types
✓	✓	✓	EndoPP	69	23
✓	×	×	EndoNN	42	14
×	✓	✓	ExoPP	69	23
×	×	×	ExoNN	3	1
×	×	✓	ExoNP	39	13
×	✓	×	ExoPN	6	2
				228	76

Notes: the number of B-types corresponds to the number of independent observations in each treatment.

We let 'P' and 'N' denote 'Punishment' and 'No punishment', respectively. Our treatment conditions are described by whether the majority vote was considered or not (Endo or Exo), whether the majority voted for the modified PGG or the base PGG (the first 'P' or 'N' after Endo or Exo), and whether the modified PGG or the base PGG was actually implemented (the second 'P' or 'N'). In ExoNP, for example, the majority voted for playing the base PGG, their vote was not considered and it was randomly determined that the modified PGG would be implemented. An intended consequence of our design is a very low number of observations in ExoNN and ExoPN, which are therefore not analyzed. This also implies that we do not analyze the data of EndoNN, as the relevant treatment comparison would be ExoNN. Analyzing the data of EndoNN in isolation does not contribute to our understanding of sanctioning institutions as none is implemented.

4.2.3 Procedures

The computerized experiment was conducted at the BonnEconLab of Bonn University. Subjects were recruited on-line with hroot (Bock, Baetge, and Nicklisch, 2014), while the software implementation was done with z-Tree (Fischbacher, 2007). A typical session lasted approximately 60 minutes and the average earnings were 13.25 Euro, including a 2 Euro show-up fee. A total of 324 subjects participated in 13 sessions (11 sessions with 24 subjects and 2 sessions with 20 subjects).[9] In order to keep instructions neutral the base PGG and the modified PGG were called "Version 1 (without deduction points)" and "Version 2 (with deduction points)", respectively.[10] In order to ensure subjects' understanding of the instructions a set of control questions was administered before the start of part 1 and another set of control questions before the start of part 2. Both parts only started when all subjects had answered them correctly. Feedback on payment (from two randomly picked periods, one from each part) was only provided after part 2 of the experiment. At the end of the experiment subjects were asked to fill out a questionnaire that gathered their demographic characteristics (see Appendix 4.6.2).

4.3 Predictions

In this section we draw on existing empirical evidence to put forward hypotheses on treatment effects for punishment and contribution choices. Further, we provide a theoretical framework that can rationalize the predicted treatment differences. Further details of the theoretical analysis, like equilibrium predictions, can be found in Appendix 4.6.1.

[9] We ran two pilots beforehand that resulted in a very low number of groups opting for the modified PGG. They did not exclude the punishment of cooperators and one did not have the first part of the experiment. The results of the pilot echo much of the literature on institutional choice in that most inexperienced subjects tend to prefer the simpler environment of the base PGG (see the discussion in Section 4.1 and footnote 5). We therefore added part 1 to the experiment.

[10] Appendix 4.6.2 contains a translation of the original German instructions.

4.3.1 Treatment Effects

Centralized formal punishment institutions are found to be more effective in enforcing cooperative behavior in social dilemmas like the linear public good game (Tyran and Feld, 2006, Kamei, 2016) or the prisoner's dilemma (Dal Bó, Foster, and Putterman, 2010) when they are endogenous. This means that for a given amount of punishment, cooperation is higher when the punishment institution was implemented as the outcome of a majority vote rather than through an exogenous process. Put differently, an endogenous formal sanction is more effective in enhancing cooperation than its exogenous counterpart. In Dal Bó, Foster, and Putterman, 2010 this difference can be ascribed to the fact that in the endogenous setting the institutional outcome is the result of a voting process, which is not the case in the exogenous setting. A sanctioning institution selected through majority voting may be perceived as more legitimate and can therefore trigger higher compliance vis-à-vis an exogenous institution. In particular, a direct and causal link between the voting outcome and the adopted institutional setting is crucial for high compliance. Whenever this link is severed, as when institutions are adopted exogenously, we can expect individuals to comply less.

Several empirical studies show that uninvolved third-parties are willing to sacrifice part of their own income to sanction non-cooperative behavior, both in one-shot and in repeated interaction (Almenberg et al., 2010; Engel and Zhurakhovska, 2013; Fehr and Fischbacher, 2004a,b; Henrich et al., 2006; Kurzban, DeScioli, and O'Brien, 2007; Nikiforakis and Mitchell, 2014). Given that in our experiment the B-types receive identical information about the vote outcome in the ExoPP and EndoPP treatments, they should hold similar beliefs on the cooperative disposition of the A-types and choose similar punishment vectors. If punishers however anticipate the positive effect of participating in the implementation process on perceived legitimacy and cooperation, those in EndoPP may believe that a given punishment level is more likely to turn a defector into a cooperator in EndoPP than in ExoPP. Consequently, they would require less punishment points to reach a certain cooperation level among the A-types when the institutional process is endogenous, rather than exogenous,

and may therefore choose lower punishment in EndoPP.[11]

Hypothesis 4.1.
Punishers anticipate that punishment is more effective in enhancing cooperation when the punishment institution is endogenously implemented and therefore choose lower punishment than when the implementation is exogenous.

In the first period of part 2 contribution decisions are yet unaffected by the implemented punishment. Controlling for the A-types' beliefs about the punishment decisions, the empirical evidence suggests that higher contributions to the public good in the first period of the game should be observed in EndoPP as compared to ExoPP. This is due to the previously discussed endogeneity premium on cooperation.

Hypothesis 4.2.
Controlling for the A-types' beliefs about the punishment vector, cooperation in the first period is higher if the punishment institution is endogeneously implemented (endogeneity premium).

From the second period onwards, public good provision may depend on the extent of punishment assigned in each treatment. The harsher the implemented punishment the more subjects might contribute to the public good (see Section 4.3.2 for further explanation). It is ex ante unclear how the positive effect of endogeneity on cooperation will balance out with its presumably negative effect on punishment.

4.3.2 Theoretical Framework

We put forward a theoretical framework to illustrate how endogeneity may affect third-party punishment via legitimacy and effectiveness concerns. For parsimony, the analysis is restricted to a simplified stage game. In the first stage A-types choose between the base PGG and the modified PGG through majority voting. In case the modified PGG is implemented, the B-type decides

[11] An alternative mechanism through which the punishment decision of the B-type may be influenced is his perceived obligation to enforce the cooperation rule by reducing the defectors' incomes. The punisher may be more willing to punish knowing that the voting of those he rules over was decisive for him being in that position, while under the exogenous institution punishers may feel less bound to spend income on punishment.

on a punishment vector, which specifies how many points should be deducted from defecting players for each possible number of defectors. The punishment vector is then revealed to the A-types, who subsequently make their contribution decisions. If the base PGG is implemented the B-type do not have the option to punish and the A-types simply interact in the PGG. An equilibrium analysis of this stage game can be found in Appendix 4.6.1. In this section we use the theoretical framework to explain how third-party punishment may affect A-types' contributions depending on the institutional process that introduces punishment.

We employ the outcome-based social preference model of Charness and Rabin, 2002 ('CR preferences' henceforth). This model posits that individuals care not only about their own payoff but also about the payoff of the worst-off individual and the sum of payoffs in their group. That is, CR preferences incorporate both Rawlsian (or minimax) and efficiency (or utilitarian) concerns. The fact that the outcome-based version of CR preferences takes efficiency gains into account is important, as cooperation substantially increases social surplus in our setting.[12] CR preferences for an A type subject are expressed by:

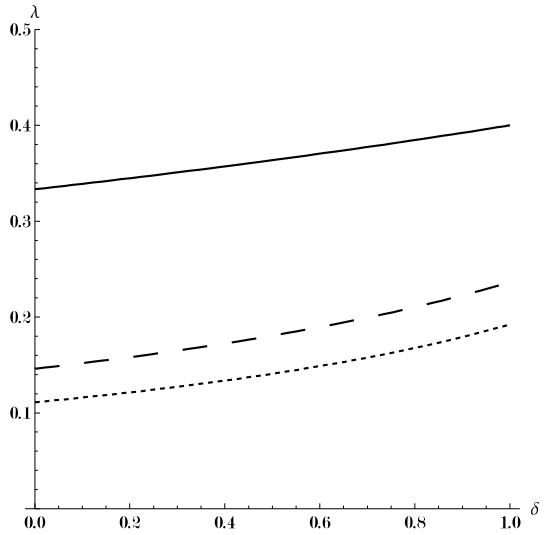
$$U_{A_i}(\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B) = \\ (1-\lambda)(\pi_{A_i}) + \lambda\Big[\delta \min[\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B] + (1-\delta)(\pi_{A_1} + \pi_{A_2} + \pi_{A_3} + \pi_B)\Big] \tag{4.4}$$

where π_{A_i} and π_B are the payoffs of the A-types and the B-type as defined in equations 4.2 and 4.3, with $i = 1, 2, 3$ indexing the three A-types in the group. $\lambda \in [0, 1]$ measures how much individual i cares about the welfare of the other subjects he is matched with, and $\delta \in [0, 1]$ governs individual i's trade-off between the payoff of the worst-off individual and the maximization of social surplus. Standard preferences are nested in the model ($\lambda = 0$), but we restrict our attention to the case of $\lambda > 0$. The CR utility function for the B-type is defined accordingly.

[12]The descriptive and predictive content of the Charness and Rabin model has been assessed in laboratory experiments (e.g., Daruvala, 2010; Engelmann and Strobel, 2004a) and it fares well when compared to other social preference models. The model has been used to derive theoretical predictions in experimental works close to ours (e.g., Markussen, Putterman, and Tyran, 2014; Sutter, Haigner, and Kocher, 2010).

In what follows we assume that the A-types' CR preferences are homogenous and common knowledge among A-types.[13] Unlike the standard preferences case, contributing to the public good can be an equilibrium outcome, both when a third-party punisher is absent, as in the base PGG, and when it is present, as in the modified PGG.[14] In the former case cooperation is a Nash equilibrium if enough weight is put on other individuals' welfare. The condition on λ and δ is given by the solid line in Figure 4.2 (cooperation is a Nash equilibrium above it). Regardless of δ, the condition is met if $\lambda \geq 0.4$ for the three A-types (see Appendix 4.6.1 for the derivations). To put this number in perspective, Daruvala, 2010 finds an average value of $\lambda = 0.397$ in her experimental study, which means that cooperation is an equilibrium outcome for a non-negligible fraction of the population.

Figure 4.2: Cooperation and punishment effectiveness



Notes: Cooperation is a Nash equilibrium above the depicted lines. Each line refers to a different case: the solid line is drawn for no punishment ($d = 0$), the dashed line represents the case when $d = 5$ and $e = 1$ and the dotted line represents the case when $d = 3$ and $e = 2$.

[13]Note that our results hinge on the specific assumptions we make about common knowledge and homogeneity of preferences. These simplifying assumptions allow us to clearly illustrate why treatment difference may occur.

[14]Under the assumptions of selfishness and rationality A-types do not contribute to the public good and B-types do not incur costs to punish defectors. As a consequence, A-types are indifferent between the base PGG and the modified PGG.

The possibility of being punished in case of defection in the modified PGG changes the A-types' incentives to cooperate. Recall that the B-type decides on the punishment vector $\mathbf{d} = (d_1, d_2, d_3)$ using the strategy method, where the index refers to the number of defectors. Punishment is credible in our framework: the B-type decides on a binding punishment vector that is announced to the A-types before they make their contribution decisions.[15] Given that each punishment point assigned by the B-type leads to a threefold reduction of a defector's income, the B-type's decision can substantially alter the A-types' incentives. By choosing positive punishment, the B-type may successfully deter A-types from defecting. This is explained by the fact that high realized punishment sacrifices efficiency and may decrease the minimum payoff. The threat of punishment leads most A-types with CR preferences to cooperate. The B-type implements the punishment threat because the resulting cooperation outcome increases her own utility through higher efficiency and a higher minimum payoff.

To illustrate how punishment and its degree of effectiveness influence the behavior of A-types, we restrict attention to the simplest punishment vector: $d_1 = d_2 = d_3 = d$. In this case the punishment of each A-type is the same regardless of what others do. Note that $d_m = 0$ if the A-type cooperates. The A-types take into account the punishment vector picked by the B-type when deciding to cooperate or defect. In addition to the multiplication factor $r = 3$ that applies to each punishment point, the effect of punishment points may be magnified or attenuated by the legitimacy of the punisher who assigns them (see Section 4.3.1). We refer to this as the effectiveness of punishment, and denote it by the parameter e, which is a positive constant. Incorporating punishment into the A-types' utility function leads to:

$$\begin{aligned} &U_{A_i}(\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B, \mathbf{d}, e) = \\ &(1-\lambda)(\pi_{A_i} + (1-e)rd) + \lambda\Big[\delta \min[\pi_{A_1} + (1-e)rd, \pi_{A_2} + (1-e)rd, \pi_{A_3} \\ &+ (1-e)rd, \pi_B] + (1-\delta)(\pi_{A_1} + \pi_{A_2} + \pi_{A_3} + \pi_B + m(1-e)rd)\Big] \end{aligned} \tag{4.5}$$

The effectiveness e can magnify or dampen the utility impact of punishment points. For $e = 1$ the payoff function is identical to that in Equation 4.3.2. To

[15] In most other second- or third-party punishment games punishment is not credible because there is no incentive to punish defectors ex-post, i.e., the decision to punish is taken after the public good players have made their contribution decisions.

illustrate our point we introduce two hypothetical cases: one in which punishment is high ($d = 5$) and efficiency is low ($e = 1$) and one in which punishment is low ($d = 3$) and efficiency is high ($e = 2$). The regions above the dashed and dotted lines in Figure 4.2 depict the area where cooperation is a Nash equilibrium for these two cases, respectively. Comparing the zero punishment case ($d = 0$, solid line) with the high punishment case ($d = 5$ and $e = 1$, dashed line) shows that the higher the punishment, more CR preference types can be brought to cooperate. Comparing the two punishment cases we see that despite punishment being higher in the first one, the second punishment vector brings more CR preference types to cooperate due to the higher effectiveness of punishment. All else equal, if punishment is more legitimate and therefore more effective in the EndoPP treatment, more A-types will choose to cooperate as compared to ExoPP.

Important caveats apply to this illustration, namely the ad hoc nature of the effectiveness parameter and the absence of equilibrium analysis from the perspective of B-types. In Appendix 4.6.1 we derive equilibrium predictions for $e = 1$ assuming homogenous A-type preferences. The analysis delivers two important insights. First, if A-types have CR preferences, a third-party punisher who shares those preferences has an incentive to set high punishment. The goal is to deter A-types with mild social preferences, who would defect in the absence of punishment but cooperate when punishment is in place. This punishment strategy is deterrent in the utility-space because of CR preference subjects' efficiency and minimum payoff concerns. Second, given that the mild social preference types choose to cooperate if punishment is in place, and given that this entails a higher payoff, we should expect them to vote in favor of the punishment institution. In other words, subjects with mild CR preferences will cooperate only if punishment is in place and they will consequently vote for the modified PGG. The intuition is that punishment acts as a commitment and coordination device for the mild CR preference types. Since highly cooperative types ($\lambda \geq 0.4$) cooperate regardless of the punishment policy, they are indifferent between punishment and no punishment. Selfish subjects ($\lambda < 0.01$) vote against punishment. In Appendix 4.6.1 we extend the analysis to a class of punishment vectors where deducted points can differ depending on how many A-types defect.

4.4 Results

We start our analysis by investigating voting behavior and its relation to public good provision in the first part of the experiment (Section 4.4.1). As we are mainly interested in how punishment behavior depends on the way it is put in place, in the remainder we will concentrate on those treatments in which the modified PGG is implemented (EndoPP, ExoPP and ExoNP). For the most part we will analyze behavior in ExoPP and EndoPP, the treatment conditions that offer the cleanest comparison, as in both conditions the existing punishment institution is desired by the majority of individuals. In Sub-section 4.4.2 we first categorize punishers in the two conditions with respect to their punishment vectors and compare punishment between treatments. We then analyze how beliefs about the A-types' cooperativeness generally influence punishment levels. In Sub-section 4.4.3 we compare cooperation behavior across our two main conditions and discuss efficiency implications. In Sub-section 4.4.4 we analyze how revealed institutional preferences interact with punishment and cooperation decisions. Here we consider behavior in ExoNP and ExoPP, as those treatment conditions only differ with respect to the outcome of the majority voting, i.e., whether the punishment institution is desired or not.

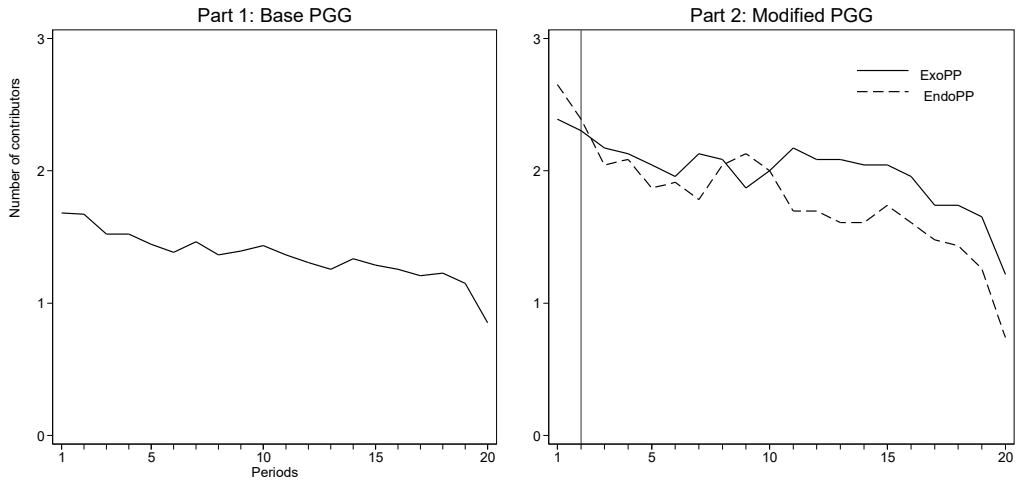
4.4.1 The Voting Decision

After interacting in the base PGG of part 1, A-types are asked to vote whether the base PGG or the modified PGG should be implemented in part 2. We find that a majority of A-types (56%) vote in favor of the modified PGG. The proportion of individuals in favor of having a non-deterrent third-party punishment institution is in line with previous comparable studies where subjects vote on the introduction of a punishment institution (see e.g., Tyran and Feld, 2006, Dal Bó, Foster, and Putterman, 2010, Markussen, Putterman, and Tyran, 2014, Kamei, 2016).[16]

[16]For example, in the experiment of Tyran and Feld, 2006 50% of subjects vote in favor of having costless non-deterrent centralized punishment. Dal Bó, Foster, and Putterman, 2010 find that around 53% prefer an environment with costless and deterrent centralized punishment. Markussen, Putterman, and Tyran, 2014 report that around 65% of subjects vote in favor of implementing costly non-deterrent centralized sanctions upon gaining experience in a social dilemma setting (around 20% of inexperienced subjects vote the same way). Kamei, 2016 finds that 42% of subjects vote in favor of a costless non-deterrent informal sanctioning institution.

The left panel of Figure 4.3 shows the average number of contributors per group in all periods of part 1. In line with the literature on repeated PGGs we observe a steady decline in public good provision over time (see e.g., Chaudhuri, 2011). The average number of contributors starts out at 1.68 and decreases to 0.85 by the end of the 20^{th} period, with a particularly pronounced drop in the last period. In order to understand how voting behavior is influ-

Figure 4.3: Number of contributors



Notes: The vertical line in the right panel indicates the point in time at which A-types are informed about the punishment vector.

enced by public good provision in part 1 of the experiment we estimate a logit model that relates voting behavior at the beginning of part 2 to the individual cooperative disposition, other group members' public good contributions and a measure of conditional cooperativeness. Following Gunnthorsdottir, Houser, and McCabe (2007), among others, we use first-period contributions as a proxy for the cooperative disposition, as the contribution decision is yet unaffected by other individuals' decisions.[17] Conditional cooperativeness of subject i is measured as the average deviation of contribution behavior in t from the other two group members' (j and k) contributions in $t-1$: $\sum_{t=2}^{20} \frac{c_{i,t} - \frac{(c_{j,t-1}+c_{k,t-1})}{2}}{19} \in [-1, 1]$. The contribution of the other group members is simply the average

[17] Gunnthorsdottir, Houser, and McCabe (2007) show in an experiment that a subject's initial contribution is an appropriate measure of their cooperative disposition.

of their contributions in part 1. Table 4.2 reports the marginal effects of the logit estimation. We find that positive experience in terms of higher average

Table 4.2: Determinants of voting decision

	(1)
First Period	0.24***
	(0.08)
Cond. Coop.	0.45**
	(0.19)
Contribution Others	-0.39***
	(0.14)
Observations	228

Notes: This table reports marginal effects of a logit regression model. The marginal effects are calculated at the means of covariates. *First Period* is a dummy variable for first-period contribution, *Cond. Coop.* is the conditional cooperativeness variable. *Contributions Others* includes the average of the other group members' contributions in part 1. Standard errors are clustered at the group level from Part 1 and indicated in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

contributions of the other group members decreases the probability to vote for the modified PGG. This is intuitive as the punishment option may not be perceived as necessary to enforce cooperation. We further find that both the first period contribution and conditional cooperativeness are significantly and positively correlated with the probability of voting for the modified PGG with punishment.

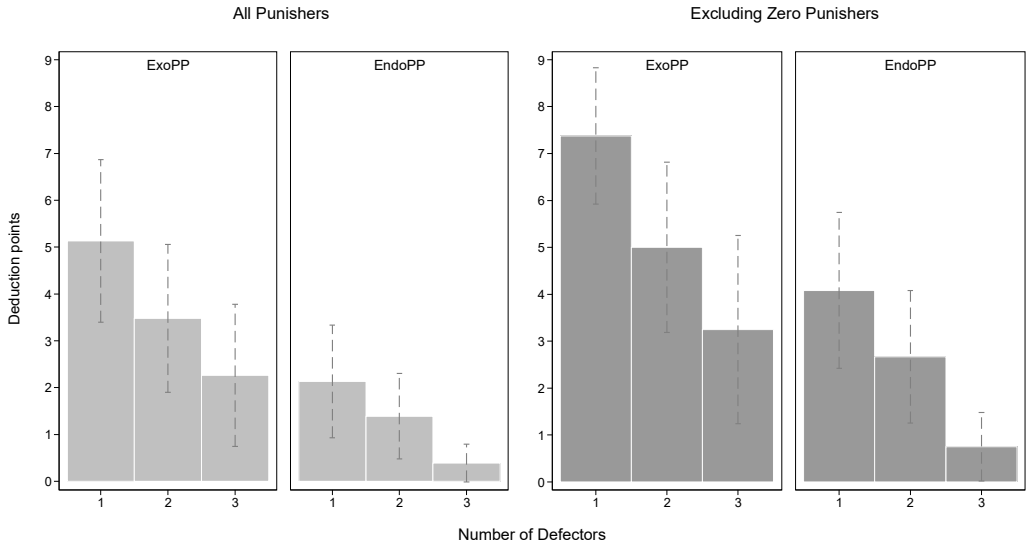
4.4.2 The Punishment Decision

Treatment differences and punisher types

Recall that the institutional preferences of groups in EndoPP and ExoPP are identical. What distinguishes them is whether the voting outcome was decisive or not. In EndoPP the punisher is endogenously appointed through majority vote, whereas in ExoPP she is exogenously appointed by a random mechanism. Figure 4.4 shows average punishment levels conditional on the number of defectors in the two treatment conditions for the complete sample

of B-types (left panel) and excluding those who assign zero punishment points in all cases (right panel). Considering all B-type decisions, punishment is on average three times higher in ExoPP as compared to EndoPP. In the former condition B-types assign an average of 3.62 points while in the latter they assign 1.30 points. In line with our hypothesis, our analysis reveals that the average of all three conditional punishment decisions is significantly higher in ExoPP than in EndoPP (two-sided Mann-Whitney test, $p = 0.03$; MW henceforth). The same is true when we consider each decision separately (MW, 3: $p = 0.07$; 2: $p = 0.07$; 1: $p = 0.01$).

Figure 4.4: Punishment decisions by B-types



Notes: The dashed lines depict 95% confidence intervals.

Result 4.1.
Punishment is significantly higher when the punishment institution is exogenously introduced as compared to when it is endogenously adopted.

We put forward a classification of third-party punishers with four categories based on their punishment vectors: zero' ($d_1 = d_2 = d_3 = 0$), 'conditional' ($d_1 \geq d_2 \geq d_3$), 'deontological' ($d_1 = d_2 = d_3$)' and 'others'. See Table 4.3 for the respective frequencies. While zero punishers are the most frequent category (39%), we find that the treatment difference is not driven by a higher number

of B-types choosing zero punishment in EndoPP. Excluding those B-types from the analysis yields average deduction points of 5.21 in ExoPP and 2.5 in EndoPP, a difference that is statistically significant (MW, $p = 0.02$). When we consider each case separately, punishment is significantly different across treatments in the one defector-case, marginally significant in the two defector-case and is insignificant in the three defector-case (MW test, 3: $p = 0.13$; 2: $p = 0.09$; 1: $p = 0.00$). The second most frequent type are conditional punishers (35%). In fact, deduction points are decreasing in the number of defectors in both treatment conditions, on average, whether we exclude zero punishers or not (Jonckheere-Terpstra test, $p < 0.01$ and $p < 0.03$, respectively). A small fraction of punishers are 'deontological' (15%), while the remaining 11% constitute the 'others' category.

Table 4.3: Type classification of punishers

	EndoPP	ExoPP	Total
Conditional	30%	39%	35%
Deontological	4%	26%	15%
Zero	48%	30%	39%
Other	17%	4%	11%
Total	23	23	46

Notes: Numbers are rounded and therefore do not necessarily sum up to 100%.

The (Expected) Effectiveness of Punishment

A possible explanation for the different punishment levels observed in our two main treatment conditions is a difference in the expected effectiveness of deduction points in increasing cooperation. If third-party punishers in EndoPP expect deduction points to be more effective due to the higher perceived legitimacy of punishment, ExoPP and EndoPP punishers may implement different punishment policies.

The B-types report their beliefs about the behavior of the A-types in part 1 and part 2 of the experiment: they are asked to indicate in how many periods they think 0, 1, 2 or 3 A-types will contribute to the group account in the 20 periods. In order to test the effectiveness conjecture we relate the change in beliefs of the B-types from part 1 to part 2 to the number of assigned deduction

points. As feedback on public good provision in both parts is only provided at the end of the experiment, beliefs about contribution rates in part 2 are unaffected by part 1 outcomes. However, the punishment vector is indicated before the belief elicitation, which means that part 2 beliefs reflect the B-type's punishment decision. The difference in beliefs represents the expected change in cooperation behavior as a result of the implemented punishment vector.[18]

Our variable of interest is the expected average change in cooperation per assigned punishment point. We compute it as the difference between a B-type's belief about the total number of contribution events (every time that $c_i = 1$) in part 2 and part 1, divided by the total number of deduction points that were assigned to A-types in part 2.[19] This variable provides a measurement of a punishment policy's expected effectiveness. Confirming our hypothesis, this measure is significantly higher in EndoPP as compared to ExoPP (MW, $p = 0.02$), taking an average value of 2.61 and 0.69 respectively. A third-party punisher in EndoPP believes that one deduction point leads to an increase of 2.61 cooperation events in part 2, taking part 1 as the reference point. In other words, third parties believe that a deduction point in the endogenous institution is 3.8 times more likely to increase cooperation than in the exogenous one. This analysis necessarily excludes zero punishers, as the effectiveness variable is not defined in the absence of assigned punishment points. Comparing the beliefs of zero punishers across the two treatments renders no statistical significance (MW, $p = 0.44$).

The questionnaire that the B-types answer after the belief elicitation in part 2 provides a further assessment of the effectiveness rationale. We elicit the B-types' expected effectiveness of punishment by asking them to report the probability that a defector who has been assigned one deduction point will change into contributing in the next round assuming the other two group members cooperated in the current round. In EndoPP punishers indicate a mean probability of 54%, while the corresponding percentage is 45% in ExoPP. This difference falls short of statistical significance (MW, $p = 0.30$), but underlines the higher expected effectiveness of punishment in the endogenous case. We

[18]Note that the B-types face different groups in part 1 and part 2. The changing group composition is however irrelevant for comparing changes in the punishers' beliefs between EndoPP and ExoPP, as in both treatments groups are randomly composed in part 1 and in part 2 all groups have voted with a majority in favor of the modified PGG. Other factors, like the repetition of the same game in parts 1 and 2, may also play a role in expectation formation, but these are constant across treatments.

[19]The variable has a missing value in case the B-type assigns no deduction points.

conclude that the differences in punishment can be explained by differences in the expected effectiveness of the assigned deduction points.

A related question is whether the expected effectiveness differential materializes. We can answer this question by analyzing how the A-types respond to the number of received deduction points in part 2. Table 4.4 presents the estimation results of a panel model where the dependent variable is a dummy that takes the value 1 if an individual increases the contribution from the previous to the current period, and 0 otherwise. The explanatory variables are the number of received deduction points in the previous period, a treatment dummy, the interaction of the latter two variables and the other group members' average contributions in the previous period. We exclude the first period's contribution decision as subjects learned the punishment vector in the second period only. The results show that received deduction points are

Table 4.4: Punishment effectiveness and cooperation

	(1)
Endo	0.58
	(0.47)
Punishment_{t-1}	0.55***
	(0.07)
Endo*Punishment_{t-1}	0.06**
	(0.03)
Contribution Others_{t-1}	-0.65***
	(0.15)
Observations	2484
Number of Groups	46
Number of Subjects	138

Notes: This table reports marginal effects calculated at the means of covariates using a logit panel model with mixed effects (including random effects at the subject level and Part 2 group level). *Endo* is a treatment dummy variable taking the value of 1 for EndoPP and 0 for ExoPP. $Punishment_{t-1}$ takes the number of received deduction points in $t-1$, $ContributionOthers_{t-1}$ takes the value of other group members' average contribution in $t-1$. The remaining variables are interaction terms. Interaction effects are calculated by the procedure proposed in Ai and Norton (2003) and Norton, Wang, and Ai (2004). Standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

associated with a significant increase in next period's contribution. The in-

teraction effect between the treatment variable and the number of deduction points is positive and significant, which means that a given amount of deduction points has a more pronounced effect on switching to cooperation when the sanctioning institution is endogenous. The fact that the B-type's punishment policy is a direct consequence of a majority decision leads her to assign fewer deduction points, which proves to be more effective vis-*à*-vis the deduction points assigned by a B-type who is appointed by chance.[20]

Result 4.2.
In line with the third-party punishers' beliefs, realized punishment is more effective in increasing cooperation among A-types when it is endogenously adopted than when the punishment institution is exogenously introduced.

Cooperation beliefs

As discussed in Section 4.3.2 punishment may incentivize mildly cooperative subjects to contribute to the public good, whereas selfish subjects will not be affected by any punishment decision. B-types should therefore only be willing to incur punishment costs if they believe that A-types are mildly cooperative and thus susceptible to respond to punishment. We use the B-types' beliefs about the cooperation behavior of the A-types in part 1 as a proxy for a general belief about the A-types' cooperativeness and investigate how they relate to the punishment decision. We therefore estimate a regression model with the B-types' average number of deduction points as dependent variable. The independent variables are the beliefs about the number of cooperation events in part 1, a dummy for the majority vote outcome and a dummy that indicates whether the vote was overruled. Our analysis includes observations from the treatment conditions EndoPP, ExoPP and ExoNP. Estimation coefficients are presented in Table 4.5. We observe that cooperativeness is positively related to higher punishment. This is in line with the idea that very selfish A-types are expected to be unresponsive to punishment, and B-types therefore do not want to waste costly punishment on them. If A-types are very cooperative there is less need for punishment, whereas if they are mildly cooperative the introduction of punishment provides the right incentives for cooperation. Furthermore,

[20]The results and significance levels are robust to including a random effect at the session level. Results are also qualitatively similar to a model in which the dependent variable is the first difference of contributions and robust to the inclusion of an interaction term between others' contributions and punishment points. These results are available upon request.

the regression reveals that punishers in the overruled conditions (ExoPP and ExoNP) assign significantly higher average punishment than those in EndoPP, which confirms the non-parametric result from the comparison of ExoPP and EndoPP. The significant positive coefficient for a majority vote in favor of punishment indicates that punishers behave in line with the A-types' majority will (see Section 4.4.4 for a detailed discussion on the role of institutional preferences).

Table 4.5: Punishment decision and individual cooperation beliefs

	(1)
Expected Cooperation	0.04**
	(0.02)
Majority Vote	1.79**
	(0.88)
Overrule	1.77**
	(0.76)
Constant	-1.97*
	(1.15)
Observations	59
R-squared	0.190

Notes: Least squares regression. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

4.4.3 Public Good Provision and Efficiency

Having identified significant differences in punishment between EndoPP and ExoPP we now look at the A-types' contribution behavior in part 2 of the experiment. The average number of contributors in each of the 20 periods is depicted in the right panel of Figure 4.3.

We start by investigating first-period differences in contributions across the two treatments. Since the punishment vector is unknown at this point, only institutional differences and the A types' beliefs about the punishment vector may affect their contribution behavior. Table 4.6 presents the marginal effects of a logit regression with first-period contributions as dependent variable and the A-types' punishment beliefs and a treatment dummy as independent

variables. We find that the probability to contribute to the public good in the first period is higher in Endo with marginal significance. This finding is in line with the existing literature on the cooperation-enhancing effect of endogenous institutions (see Dal Bó, Foster, and Putterman, 2010) as captured by our hypothesis. It is not only the information implied in the voting decision that affects cooperative behavior, but it matters whether an institution that may punish non-cooperative behavior is exogenously imposed or chosen by the affected individuals themselves.

Table 4.6: First-period contribution determinants

	(1)
Endo	0.11*
	(0.06)
Belief 3 Defectors	-0.04***
	(0.01)
Belief 2 Defectors	0.02
	(0.03)
Belief 1 Defector	0.01
	(0.02)
Observations	138

Notes: Logit model. Reported results are marginal effects calculated at the means of the covariates. Standard errors are clustered at the Part 1 group level and indicated in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1. Endo is a treatment dummy variable taking the value of 1 for EndoPP and 0 for ExoPP. BeliefDeduction'x' are A-types' punishment beliefs for the case of 'x' defectors.

We have shown that cooperation in EndoPP is significantly higher in the first period. After the first period, punishment decisions are revealed and implemented and may influence subsequent contribution behavior. We find no statistical significance between ExoPP and EndoPP contributions if we consider the entire 20 periods (MW, $p = 0.24$). Singling out contributions in the first 10 periods also does not produce a statistically significant difference (MW, $p = 0.85$). Cooperation levels in the two treatments seem to converge after an initial difference, possibly due to the higher punishment implemented in ExoPP. In fact, in the last 10 periods there are on average more A-types contributing to the group account in ExoPP, i.e., when the punishment institutions has been exogenously introduced as compared to endogenously adopted.

This difference is marginally significant (MW, $p = 0.09$).

While the exogenous institution slightly outperforms the endogenous one in sustaining cooperation, this is done at the expense of higher punishment. An efficiency assessment of endogenous and exogenous institutions must take this into account. The punishment points received by the A-types in the two treatment conditions are depicted in the left panel of Figure 4.5. The right panel shows the average group payoff (our efficiency measure), which takes into account the punishment points deducted from the A-types and the punishment costs deducted from the B-types' endowments. We observe that punishment is higher in ExoPP, in particular towards the end of part 2, which brings the earnings in ExoPP very close to those in EndoPP. For neither player type we observe significant differences in payoffs between treatments (MW test, A: $p = 0.39$; B: $p = 0.63$).[21][22]

Result 4.3.
We find evidence for an endogeneity premium: first-period cooperation rates are higher when the punishment institution is endogenous. Overall, cooperation levels and efficiency are independent of the institution-generating process.

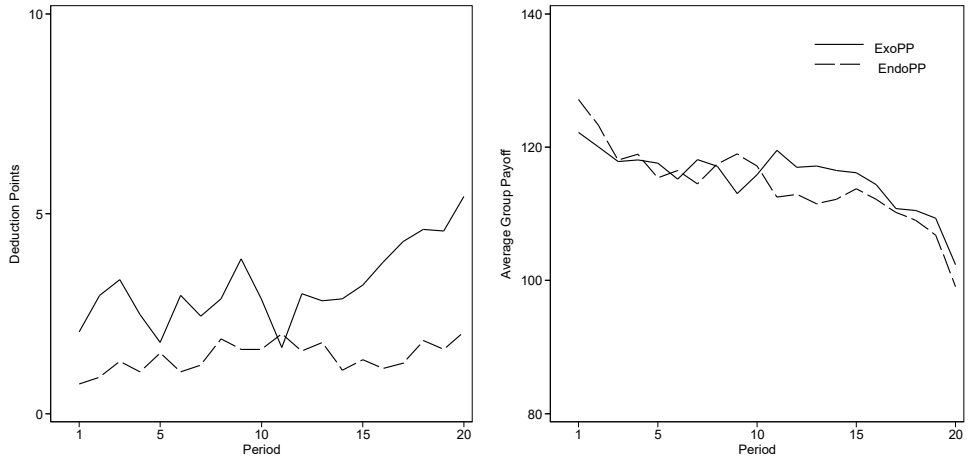
4.4.4 The Role of Institutional Preferences

The voting outcome is public information for all subjects in a group. In order to analyze whether the punishment and cooperation decisions are influenced by the majority decision of the A-types we look at punishment decisions in the Exo treatments that differ with respect to the outcome of the majority vote: ExoNP and ExoPP. We find that average punishment is significantly lower in ExoNP as compared to ExoPP with 1.44 and 3.62 points respectively (MW test, $p = 0.07$). The regression in Table 4.5 echoes this finding. In the cases of one, two and three defectors, the respective average punishment levels are

[21] There are no significant differences in payoffs between treatments in the first 10 or the last 10 periods (MW test, periods 1-10, A: $p = 0.90$; B: $p = 0.64$, periods 12-20, A: $p = 0.16$; B: $p = 0.75$). Our independent observation for these tests is the average payoff in a group for the A-types and the B-types respectively.

[22] The existence of the punishment institution per se does not have efficiency implications: comparing payoffs for groups where average punishment is zero we find that differences in payoffs for the A-types are insignificant (MW test, $p = 1$). The B-types' payoffs do not differ across treatments as they keep their fixed endowment and do not spend money on deduction points.

Figure 4.5: Punishment and efficiency across treatments



Notes: Average group payoff is defined as the average payoff of the three A-types and the B-type in each group.

1.92, 1.54 and 0.85 in ExoNP and 5.13, 3.48 and 2.26 in ExoPP. The separate case-specific analysis for different numbers of defectors yields that only the difference for the one-defector case is statistically significant (3: $p = 0.40$, 2: $p = 0.15$, 1: $p = 0.04$).[23] Our results suggest that the votes of the A-types matter for the severity of the punishment that is implemented. In the analysis of Section 4.4.1 we find that selfish A-types are more likely to vote against punishment. Thus, the vote outcome gives an indication for the cooperativeness of the A-types. In fact we find that B-types have more pessimistic beliefs about A-types' contributions in part 2 in ExoNP compared to ExoPP (average number of contribution events in ExoNP 38.77 and in ExoPP 47.48). This difference, however, falls short of statistical significance ($p = 0.13$). B-types may choose a lower punishment in ExoNP since punishment is unlikely to deter A-types and the money spent on punishment would therefore be wasted. This is reflected by the substantially larger share of zero punishers in ExoNP

[23]While the B-types' behavior is responsive to the process (Exo vs. Endo) as well as to the majority vote outcome, this is not anticipated by the A-types. Comparing the A-types' beliefs about the average punishment of the B-types between ExoPP and ExoNP as well as between ExoPP and EndoPP reveals that A-types do not anticipate the B-types' consideration of the voting outcome (MW test, $p = 0.88$) or the effect of endogeneity on punishment (MW, $p = 0.64$).

(54%) compared to ExoPP (30%). In addition, the punisher may simply want to respect the majority's will (no punishment) and therefore assigns few punishment points.

Comparing public good contributions between ExoNP and ExoPP reveals that contributions are significantly higher when the majority voted in favor of implementing the modified PGG with punishment (MW tests, $p \leq 0.04$), which can be explained by the selection of cooperative subjects into ExoPP.[24]

4.5 Conclusion

This paper investigates the role of institutional endogeneity on third-party sanctioning and the resulting consequences for cooperation behavior. A growing experimental literature on institutional choice has documented the existence of an endogeneity premium on cooperation when formal sanctioning institutions are selected through a democratic procedure. That is, groups that can choose the sanctioning institutions under which they interact tend to cooperate more. It has been shown that this phenomenon is due to the participation rights granted to groups, and not to self-selection into a preferred institution or signaling of a willingness to cooperate.

Our study compares the behavior of third-party punishers who are elected by the group she is supposed to sanction to that of third-party punishers who are appointed by chance. We show that for third-party punishment institutions endogeneity leads to milder sanctions. This result can be explained by the higher effectiveness of punishment in changing defectors' behavior in the endogenous case. A third-party punisher that is endogenously appointed anticipates that her sanctions are effective in turning defectors into cooperators, and therefore a lower level of punishment is deemed necessary. When the same institution is imposed exogenously punishment tends to be harsher but is not more empowered in enhancing cooperation. In spite of endogenous sanctions initially leading to more cooperation, overall the two environments exhibit identical outcomes, both in terms of cooperation and efficiency.

The idea that third-party punishers may be more lenient when an institution is endogenous to the affected individuals has previously been suggested

[24]Our units of observations here are the average number of contributions within a group in the first 10 and last 10 periods respectively.

by Feld and Frey (2002). The authors find that in cantons that score higher on a general direct democracy index (Stutzer, 1999) tax authorities impose lower maximum fines for tax evasion and lower fines in the case of self-denunciations. These implications are drawn from a context in which - unlike our setting - there is no direct link between the democratic participation rights and the task of the third-party (the tax authority), and selection and signaling effects are present. Their result resonates with our finding of lower punishment levels for defection in our endogenous institutional setting.

We can draw two main implications. First, externalities of a democratic process need to be considered when designing institutions, as individuals outside the decision process may be influenced by it. In our particular case, with an endogenous process third parties choose milder sanctions that are as effective with respect to overall efficiency as higher sanctions in the exogenous case. Second, applying an endogenous implementation process to punishment institutions may be particularly useful when punishment costs are high, as lower punishment is required to enhance cooperative behavior than when an exogenous process is applied. On the other hand, one needs to take into account that exogenous institutions may outperform their endogenous counterparts with respect to cooperation levels, as generally higher punishment levels are established. It is therefore crucial to ponder the effects that the implementation process of institutions has on the different variable features of the institutional design.

4.6 Appendix

4.6.1 Model Predictions

In what follows we assume that both A-types and B-types have CR preferences as defined in Section 4.3.2. Preferences of A-types are homogenous and common knowledge to the A-types. The B-type is aware of the preference homogeneity of the A-types, but does not know the exact values of δ and λ.

Part 1

We first look at the PGG without punishment. Given that the B-type is neither influenced nor receives information on what the A-types do, we do not consider her as a player of this game. The 3 A-types, indexed as A_i (with i={1,2,3}), have to make their contribution decision in private, which consists of allocating their endowment E_A to either the group or the private account ($c_i = 1$ and $c_i = 0$, respectively). We will refer to these decisions as cooperation versus defection or contributing versus not contributing. The utility of an A-player is defined as:

$$U_{A_i}(\pi_{A_1}, \pi_{A_2}, \pi_{A_3}) = (1-\lambda)\pi_{A_i} + \lambda[\delta \min[\pi_{A_1}, \pi_{A_2}, \pi_{A_3}] + (1-\delta)(\pi_{A_1} + \pi_{A_2} + \pi_{A_3})]$$

$$\text{with} \quad \pi_{A_i} = E_A(1 - c_i + \alpha G)$$

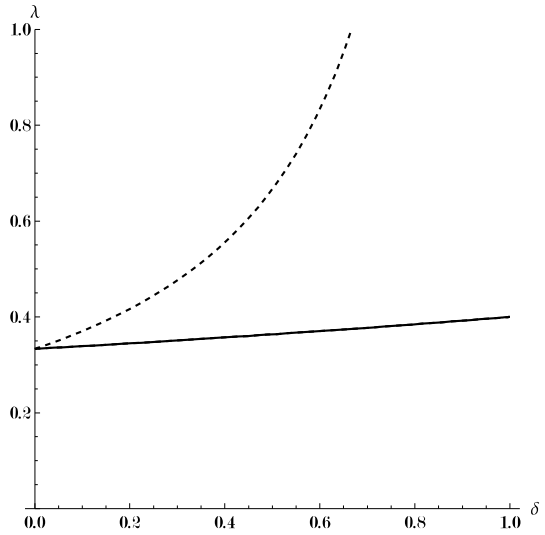
Proposition 4.1.
For $\lambda \geq \frac{1-\alpha}{2\alpha(1-\delta)}, c_i = 1$ is a dominant strategy and full cooperation is the unique Nash equilibrium. For $\frac{1-\alpha}{2\alpha(1-\delta)} > \lambda \geq \frac{1-\alpha}{2\alpha(1-\delta)+\delta}$ both full cooperation and full defection are Nash equilibria and a mixed strategy equilibrium exists. For $\lambda < \frac{1-\alpha}{2\alpha(1-\delta)+\delta}, c_i = 0$ is a dominant strategy and full defection is the unique Nash equilibrium.

Proof. Define $U_{A_i}(c_i, c_{-i}; \lambda, \delta)$ as the utility obtained from the material payoffs associated with the contribution decision profile (c_i, c_{-i}) and parameters λ and δ. Cooperation is preferred to defection if $U_{A_i}(c_i = 1, c_{-i}; \lambda, \delta) \geq U_{A_i}(c_i = 0, c_{-i}; \lambda, \delta)$. This implies $\lambda \geq \frac{1-\alpha}{2\alpha(1-\delta)}$ when two other players defect, and $\lambda \geq \frac{1-\alpha}{2\alpha(1-\delta)+\delta}$ both when one other player cooperates and when two other players cooperate. When cooperation (defection) is a best response the corresponding Nash equilibrium follows. For values of λ such that $\frac{1-\alpha}{2\alpha(1-\delta)} > \lambda \geq \frac{1-\alpha}{2\alpha(1-\delta)+\delta}$, the

players cooperate if one or two others do the same, but defect when the other two defect. Full cooperation and full defection are both a Nash equilibrium, and a mixed strategy equilibrium exists. □

For the MPCR used in the experiment, $\alpha = 0.6$, the above conditions simplify to $\lambda \geq \frac{1}{3(1-\delta)}$ and $\lambda \geq \frac{2}{6-\delta}$. Figure 4.6 depicts these conditions.

Figure 4.6: Preference parameters and cooperation



Notes: The solid (dashed) line represents the second (first) condition set forth in Proposition 4.1. In the region above the solid line cooperation is a Nash equilibrium. It is unique above the dashed line and payoff-dominant otherwise. Below the solid line defection is the unique Nash equilibrium.

Part 2

We now turn to the analysis of the stage game when A-types can choose between the base PGG and the modified PGG. In the first stage A-types choose between the base PGG and the modified PGG through majority voting. In case the modified PGG is selected, the B-type decides on a punishment vector, which specifies how many points should be deducted from defecting players for each possible number of defectors. The punishment vector is then revealed to the A-types, who subsequently make their contribution decisions. If the base

PGG was chosen no punishment option for the B-type exists and the A-types simply make their contribution decisions. We mainly focus on the case that the modified PGG is implemented and start by deriving the optimal contribution decision of the A-types given the punishment vector, and then continue with the optimal punishment decision of the B-type given the vote outcome. The game is solved by backward induction.

We define the punishment vector as $\mathbf{d} = (d_1, d_2, d_3)$, where the index indicates the number of A-types that defect. Recall that the B-type cannot discriminate between defectors in a given situation, and that it is not possible to punish cooperators. The punishment points are multiplied by a factor r before being deducted from an A-type's payoff.

A-type Contributions: Since the A-types' decisions have payoff consequences for the B-type, their preferences must explicitly incorporate her welfare. The utility function becomes:

$$U_{A_i}(\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B, \mathbf{d}, e) =$$
$$(1-\lambda)(\pi_{A_i}) + \lambda\Big[\delta \min[\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B] + (1-\delta)(\pi_{A_1} + \pi_{A_2} + \pi_{A_3} + \pi_B)\Big]$$

$$\text{with} \quad \pi_{A_i} = \begin{cases} \alpha G E_A, & \text{if } c_i = 1 \\ (1+\alpha G)E_A - 3d_m, & \text{if } c_i = 0 \end{cases} \qquad \text{and} \quad \pi_B = E_B - m d_m$$

where m indicates the number of A-types that defect. For simplicity, we set $e = 1$ here. As discussed in Section 4.3, the parameter e is a measure for the effectiveness of punishment in increasing cooperation. We assume that e mirrors the perceived legitimacy of the punisher. The B-type has CR preferences identical to those of the A-types; his utility function is defined accordingly.

As in the previous sub-appendix, we start with the derivation of the A-types' best-response behavior. If two other players defect, an A-type will contribute if $\lambda \geq \frac{(1-\alpha)E_A - r d_3}{(1-\delta)[2\alpha E_A + 2(1+r)(d_3-d_2)+d_3]}$. If one other player cooperates and one other defects, or two others cooperate, an A-type player will contribute if, respectively $\lambda \geq \frac{(1-\alpha)E_A - r d_2}{\delta E_A - r d_2 + (1-\delta)[2\alpha E_A + (1+r)(2d_2-d_1)]}$ and $\lambda \geq \frac{(1-\alpha)E_A - r d_1}{E_A - r d_1 + (1-\delta)[(2\alpha-1)E_A + (1+r)d_1]}$.

For the parameter values used in the experiment ($E_A = 70$ and $r = 3$) these conditions become:

$$\lambda \geq \frac{28 - 3d_3}{(1-\delta)(84 + 9d_3 - 8d_2)} \tag{4.6}$$

$$\lambda \geq \frac{28 - 3d_2}{14(6-\delta) + (5-8\delta)d_2 - 4(1-\delta)d_1} \tag{4.7}$$

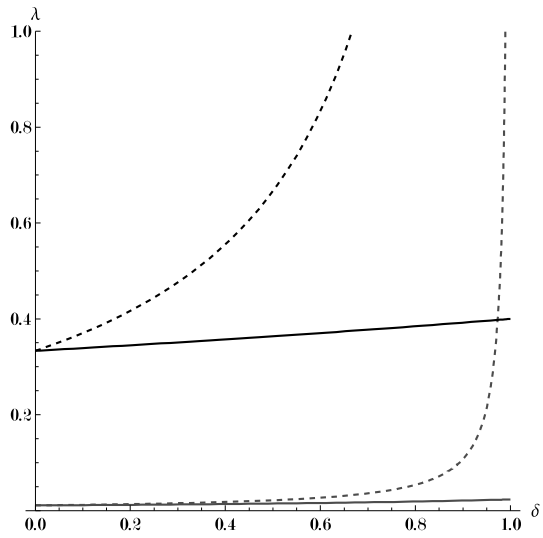
$$\lambda \geq \frac{28 - 3d_1}{14(6-\delta) + (1-4\delta)d_1} \tag{4.8}$$

For $\mathbf{d} = (0, 0, 0)$ these conditions boil down to the predictions for the base PGG. The equilibrium or equilibria that result will depend on both the parameters λ and δ and the punishment vector $\mathbf{d}$. In fact, any symmetric strategy profile is an equilibrium for some combination of parameter values and punishment vector.

Punishment and Voting: In order to make the analysis tractable we restrict the analysis to two types of punishment vectors that represent 82% of the non-zero punishment vectors in our sample. We will start with the simplest case: $d_1 = d_2 = d_3 = d$. This would correspond to a situation in which the B-type chooses the same level of punishment regardless of the A-types' behavior, which we refer to as a 'deontological' punishment vector in the main text. Imposing the same level of punishment for each possible outcome makes equations 4.7 and 4.8 identical, and renders equation 4.6 more binding than the former two as long as $d < 70/3$, which is true in our case as $d_i \in \{0, ..., 9\}$. The range of (λ, δ) for which cooperation is a Nash equilibrium is monotonically increasing in d. Figure 4.7 illustrates the point by plotting the equilibrium conditions for $d = 0$ and $d = 9$. The intuition is simple: increasing punishment renders cooperation a best-response for a wider range of CR preference types, as the dis-utility caused by punishment through efficiency and concerns for the lowest payoff increases.

For the remaining derivations we apply payoff dominance as an equilibrium selection criterion (Harsanyi and Selten, 1988). For example, we assume that for parameter configurations for which full cooperation and full defection are both a Nash equilibrium A-types will play the full cooperation equilibrium. The material payoff of cooperation is substantially higher than the one for defection: each A-type receives 126, compared to 70 in case of full defection. The latter payoff will be lower if punishment is positive. In the CR-utility space these differences will be more pronounced because of efficiency concerns. In addition, the fact that subjects vote in favor of punishment provides a strong signal towards coordinating on cooperation, in case both cooperation and defection are Nash equilibria. The B-type's utility function is defined as:

Figure 4.7: Punishment levels and cooperation



Notes: The black (red) lines represent the equilibrium conditions for $d = 0$ ($d = 9$).

$$U_B(\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B, \mathbf{d}, e) =$$
$$(1-\lambda)(\pi_B) + \lambda\Big[\delta \min[\pi_{A_1}, \pi_{A_2}, \pi_{A_3}, \pi_B] + (1-\delta)(\pi_{A_1} + \pi_{A_2} + \pi_{A_3} + \pi_B)\Big]$$

Proposition 4.2.
If the B-type is assumed to choose the same level of punishment in all cases ($d_1 = d_2 = d_3 = d$), the equilibrium is characterized by the B-type setting $d^* = 9$, selfish and mostly selfish A-types ($\lambda < \frac{1}{(93-50\delta)}$) voting against punishment and all others voting in favor of punishment.

Proof. Let $U_B(c_1, c_2, c_3, \mathbf{d}; \lambda^B, \delta^B)$ be the utility accruing to the B-type when the A-types play (c_1, c_2, c_3). She picks the punishment level $\mathbf{d} = (d, d, d)$ and has CR preferences described by λ^B and δ^B. The B-type knows that the A-types will cooperate if $\lambda \geq \frac{28-3d}{14(6-\delta)+(1-4\delta)d}$ and defect otherwise. Increasing d makes this condition less binding, i.e., full cooperation will be a Nash equilibrium for a broader range of CR preferences. We can show that

$U_B(1,1,1,\mathbf{d};\lambda^B,\delta^B) \geq U_B(0,0,0,\mathbf{d};\lambda^B,\delta^B)$ for all $(\lambda^B,\delta^B) \in [0,1]$:

$$153 + 378\lambda^B - 405\lambda^B\delta^B \geq 153 + 210\lambda^B - 293\lambda^B\delta^B \tag{4.9}$$

$$\Rightarrow \delta^B \leq 1.5 \tag{4.10}$$

As a result, the B-type will set $d = 9$ in order to make as many CR preference types as possible cooperate. Those A-types who would not cooperate for $d = 0$ but cooperate for $d = 9$, i.e., those for whom $\frac{2}{6-\delta} > \lambda \geq \frac{1}{93-50\delta}$, are better off in the latter case as cooperation entails a higher payoff $(70 + 293\lambda(1 - \delta)$ and $126 + 405\lambda(1 - \delta)$ for full defection and cooperation, respectively). Punishment does not decrease payoffs as it is merely deterrent. These A-types will thus vote in favor of punishment. A-types which are not deterred by the maximum punishment level ($\lambda < \frac{1}{93-50\delta}$) will vote against punishment. Highly cooperative A-types ($\lambda \geq \frac{2}{6-\delta}$) are indifferent between punishment and no punishment as they will cooperate in either case, and therefore are weakly in favor of punishment. □

In sum, the B-type implements a credible punishment vector that sorts A-types into their preferred institution through voting. A-types that are deterred by this punishment policy are better off under a punishment regime. Those who would be worse off under punishment do not vote for punishment.

Next, we investigate the optimal punishment vector and voting behavior under less restrictive conditions assuming a 'conditional' punishment policy. We assume that $d_1 \geq d_2 \geq d_3$, i.e., a single defector is never punished less harshly than two defectors.[25]

Proposition 4.3.
If the B-type is assumed to choose $\mathbf{d}$ *such that* $d_1 \geq d_2 \geq d_3$*, the equilibrium is characterized by the B-type setting* $d_1^* = 9$*, selfish and mostly selfish A-types* ($\lambda < \frac{1}{(93-50\delta)}$) *voting against punishment and all others voting in favor of punishment.*

[25]Such a punishment policy may exist if free-riding is deemed more deserving of punishment when at least one other A-type cooperates. Equivalently, such a policy can be rooted in the fact that it is easier to bring one defecting A-type to cooperate than achieving the same when all A-types defect. This punishment strategy would also be picked by a B-type who wanted to keep expenditure relatively constant across the three possible cases (recall that d_1 has to be paid once while d_3 has to be paid three times). In fact, the average punishment vector in the experiment conforms to $d_1 \geq d_2 \geq d_3$.

Proof. We start by describing the equilibria that can possibly occur in the PGG. When $d_1 \geq d_2 \geq d_3$, it can be shown that both condition 4.6 and 4.7 are more binding than 4.8, i.e., whenever one or both of the former are binding the latter necessarily is too. However, it is not guaranteed that condition 4.6 is more binding than 4.7. One of four equilibrium configurations can be observed, depending on $\mathbf{d}$ and the A-type's (λ, δ):

- if the conditions expressed in equations 4.6, 4.7 and 4.8 are all binding, cooperation is the unique equilibrium.
- if equation 4.6 is binding (and then necessarily also equation 4.8) but equation 4.7 is not, there exist two equilibria: full cooperation and one A-type cooperating and two defecting. Since the former involves a higher payoff for all players, we select it according to our payoff dominance criterion.
- if equation 4.7 is binding (and then necessarily also equation 4.8) but equation 4.6 is not, there are two equilibria: full cooperation and full defection. Since the former involves a higher payoff for all players, we select it according to our payoff dominance criterion.
- in all other cases full defection is the unique equilibrium.

Given that full cooperation is an equilibrium whenever equation 4.8 is binding, the B-type will make it the least binding possible in order to get as many CR preference types as possible to cooperate. Since equation 4.8 only depends on d_1 and its partial derivative with respect to it is negative, increasing d_1 lowers the λ for which full cooperation is an equilibrium. Therefore, the B-type will set $d_1 = 9$. The remaining vector entries can take any value as long as $d_1 \geq d_2 \geq d_3$. The voting behavior of A-types is identical to what was shown in Proposition 4.2. □

In brief, under preference homogeneity among the A-types the B-type chooses to punish defection by one player (d_1) as harshly as possible, as this guarantees the existence of a full cooperation equilibrium for the broadest CR preference parameter range. A-types who are sufficiently cooperative, i.e., who can be persuaded to cooperate by this punishment level, vote in favor of punishment. All others vote against punishment. The punishment of two and three defectors (d_2 and d_3) does not play a role in this result as long as we impose the payoff-dominance selection criterion. Relaxing this assumption (e.g.,

by allowing for mixed strategy equilibria) would allow us to say more about the second and third punishment vector entries, but a meaningful analysis would also require a detailed distribution of the B-types' beliefs on λ and δ.

4.6.2 Instructions

Paper Instructions

General Instructions for Participants

You are taking part in an economic experiment. Please read the following instructions carefully. You can earn money in this experiment. Your earnings depend on both your decisions and on the decisions of the other participants. At the end of the experiment, the total amount of money earned will be paid to you in cash. Additionally, you will receive a show-up fee of 2 Euro.
Throughout the experiment, monetary amounts are not quoted in Euro, but points. Your total earnings will thus be initially calculated in points. In the end the total amount of money earned during the experiment will be converted into Euro, where:

1 Point = 0.05 Euro

The experiment consists of two parts. You can earn money in both parts. So far you have received only the instructions of part 1. Instructions for part 2 will be handed out when part 1 is completed.

In this experiment there are two types of participants, A-participants and B-participants, who make different decisions. You will only get to know your own type shortly before the start of the experiment. The types will be randomly assigned. Please read the instructions about the decisions of the two types carefully.

All participants receive the same instructions. Hence, all participants receive the same information. **Talking is not permitted throughout the entire experiment.** Failure to comply will result in exclusion from the experiment and the loss of all earnings. If you have any questions, please address them to us: raise your hand and an experimenter will come to you.

On the following pages, the further course of the experiment is described in detail.

Information about the Procedure of Part 1 of the Experiment

The experiment consists of 20 periods. At the beginning of the experiment all participants will be randomly divided into **groups of 4 participants**, each group consisting of three A-participants and one B-participant. This group composition remains unchanged throughout the 20 periods. That is, you interact with the same three participants through all 20 periods.

At the end of the experiment, 1 of the 20 rounds will be randomly selected. The total amount of points earned in this period determines your payoff from part 1 of the experiment. You will not receive any information about your payoff before the end of part 2 of the experiment.

In every period, each of the three A-participants and the B-participant receive an endowment of 70 points and 153 points, respectively.

Each of the A-participants has to decide on how to allocate their endowment. There are two options:

- You choose the **private account**: Your endowment of 70 points will be allocated to the private account.
- You choose the **group account**: Your endowment of 70 points will be allocated to the group account.

The income of an A-participant is calculated differently according to the chosen account:

The **point income from the private account** directly corresponds to the amount of points allocated to it. If you allocate your endowment to the private account your income from the private account amounts to 70 points. If you allocate your endowment to the group account, your income from the private account amounts to 0 points. Nobody but yourself derives income from your private account.

Your **point income from the group account** does not solely depend on your decision, but also on the decisions of the other A-participants in your group. The point income from the group account corresponds to the sum of contributions to the group account by all three A-participants, multiplied by the factor 0.6. As soon as one of the A-participants (either you or a member of your group) chooses the group account, the group account increases by 70 points. Accordingly, the point income that every A-participant of the group receives increases by 70 x 0.6 = 42 points. The total point income of the group thereby increases by 3 x 42 = 126 points. Every A-participant of a group receives the same income from the group account, regardless of whether she contributed to the group account or not.

Depending on how the three A-participants decide to allocate their points, 4 different cases can occur: Table 1 illustrates the total group income depending on the number of A-participants who chose the group account (group-account-contributors), the number of A-participants who choose the private account (private-account-contributors) and whether an A-participant is a private- or a group-account-contributor herself, respectively.

Table 1: Decisions and Total Point Income of A-Participants

Case	Decisions of the A-Participants	Point Income of a Private-Account-Contributor	Point Income of a Group-Account-Contributor
1	3 A-participants choose the private account and 0 A-participants choose the group account	70	-
2	2 A-participants choose the private account and 1 A-participant chooses the group account	112	42
3	1 A-participant chooses the private account and 2 A-participants choose the group account	154	84
4	0 A-participants choose the private account and 3 A-participants choose the group account	-	126

As an example, we will explain case 2 of Table 1 in more detail below.

Cass 2: *Two of the three A-participants choose the private account and one chooses the group account. Hence, 1 x 70 = 70 points are in total allocated to the group account. Every A-participant receives 70 x 0,6 = 42 points from the group account. The two private-account contributors additionally receive 70 points from their private account and thus receive a total of 42 + 70 = 112 points each.*

After every period, the A-participants receive information about their point income from both the group and the private account.

While the A-participants are making their decisions, the B-participants are asked to complete a questionnaire. The corresponding instructions will be presented on the computer screen. The total point income of a B-participant equals the endowment of 153 points in every round.

The total point income of an A-participant is calculated as follows:

Point income from the private account
\+ Point income from the group account
= Total Point Income

The total point income of a B-participant is calculated as follows:

Point endowment
= Total point income

Recall: Only one of the 20 periods will be randomly selected. The total point income in this period determines your payoff from part 1 of the experiment.

Information about the Procedure of Part 2 of the Experiment

In this part, you are assigned to the same type (A-participant or B-participant) as in the first part of the experiment. The experiment consists of 20 periods. Once again, you will be randomly divided into **groups of 4 participants**. Each group consists of 3 A-participants and 1 B-participant. Your fellow group members will not be the same as in the first part of the experiment. Instead, a new group with 3 different fellow group members is formed. This grouping remains unchanged throughout the 20 periods. That is, you interact with the same three participants for all of the 20 periods.

As in part 1, 1 of the 20 rounds will be randomly selected at the end of the experiment. The total amount of points earned in this period determines your payoff from part 2 of the experiment.

In every period, each of the three A-participants and the B-participant receive an endowment of 70 points and 153 points, respectively.

As in part 1, every A-participant has to decide on how to allocate her endowment. Before heading to these decisions a vote will take place. The three A-participants vote with which of two versions of the experiment they wish the experiment to proceed (version 1 or version 2).

Version 1 - Experiment without the option to assign deduction points

In this version, the instructions remain the same as in part 1 of the experiment. The A-participants decide how to allocate their endowment. The B-participants complete a questionnaire.

Version 2 - Experiment with the option to assign deduction points

For all A-participants, the decision on how to allocate their endowment in version 2 is exactly the same as in version 1.

Additionally, the B-participant can reduce the income of the A-participants who choose the private account by **assigning deduction points**. The B-participant can also leave the income of private-account-contributors unchanged by refraining from assigning deduction points. The B-participant cannot, however, assign deduction points to group-account-contributors.

Every deduction point that a B-participant assigns to a private-account-contributor has a **deduction value** of 3 points. That is, assigning 1 deduction point reduces the private-account-contributor's income by 3 points.

Table 2: Deduction points and deduction values

Deduction points	0	1	2	3	4	5	6	7	8	9
Deduction value	0	3	6	9	12	15	18	21	24	27

Table 2 shows an overview of the resulting deduction values for all possible quantities of deduction points (0-9). If, for instance, the B-participant assigns 3 deduction points to a private-account-contributor, this leads to a deduction value of 9 points. That is, the income of the private-account-

contributor is reduced by 9 points in this round. Accordingly, if the B-participant assigns 0 deduction points to a private-account-contributor, this leads to a deduction value of 0 points. That is, the income of the private-account-contributor remains unchanged.

To each of the private-account-contributors, a maximum of 9 deduction points can be assigned. 9 deduction points lead to a deduction value of 27 points. It is not possible to assign different numbers of deduction points to particular private-account-contributors. A B-participant can assign a maximum of 27 deduction points (= 3 private contributors * 9 deduction points).

The sum of assigned deduction points to the private-account-contributors will then be deducted from the B-participant's endowment (153 points).

Hence, deduction points indicate by how many points the income of a B-participant is reduced. Deduction values indicate by how many points the income of an A-participant is reduced.

When the B-participant decides on the deduction points for the private-account-contributors, the actual decisions of the A-participants are yet unknown. Thus, decisions on the deduction points are made for the 3 possible cases when there is at least 1 private-account-contributor, i.e. independent of the yet-unknown number of private-account-contributors. The B-participant enters the deduction points for each of the three cases in table 3, which will then be presented on the computer screen. In the fourth possible case, no deduction points can be assigned, since in this case all A-participants are group-account-contributors.

Table 3: Decisions of the B-Participants

Case	Decision of the A-Participants	Deduction Points Per Private-Account-Contributor (0 – 9)
1	3 A-participants choose the private account and 0 A-participants choose the group account	
2	2 A-participants choose the private account and 1 A-participant chooses the group account	
3	1 A-participant chooses the private account and 2 A-participants choose the group account	
4	0 A-choose the private account 3A-participants choose the group account	----

As an example, we will explain case 3 of Table 3 in more detail below.

*Case 3: One A-participant chooses the private account and two A-participants choose the group account. The private-account-contributor earns 154 points and the group-account-contributors each earn 84 points (see Table 1). If, for instance, the B-participant assigns 7 deduction points to the private-account-contributor, the B-participant's endowment of 153 points is reduced by the arising cost of 7 points (153-7=146). The private-account-contributor's income is reduced by the deduction value of 3*7=21 points to 154-21=133 points. The income of both the group-account-contributors remains unchanged (84 points).*

The B-participant's decisions on the deduction points for the 3 relevant cases apply to all of the 20 periods. In each period, the deduction points determined by the B-participant apply according to the actual number of private-account-contributors. After the first period, the A-participants receive

information about the decision on the assignment of deduction points to the private-account-contributors, which the group's B-participant made for each of the 3 cases.

After the deduction point decision, the B-participant is asked to complete a questionnaire, to be presented on the computer screen. At the end of the 20 periods, the B-participant receives information about the A-participants' point allocation, the sum of deduction points assigned to the three A-participants, and their own point income in each of the periods.

The point income of an A-participant is calculated as follows:

	Point income from the private account
+	Point income from the group account
-	Deduction value (= assigned deduction points*3)
=	Total point income

The point income of a B-participant is calculated as follows:

	Point endowment
-	Sum of assigned deduction points to private-account-contributors
=	Total point income

Recall: Only one of the 20 periods will be randomly selected. The total point income in this period determines your payoff from part 2 of the experiment.

The Vote between Version and Version 2

Before the A-participants make a decision on the allocation of points, a vote takes place. The three A-participants vote on whether they wish to proceed with version 1 (without the option to assign deduction points) or with version 2 (with the option to assign deduction points) of the experiment.

After the A-participants have cast their vote, the computer randomly determines whether the vote will be considered.

- If the computer determines the vote to be **considered**, the **majority** determines whether version 1 or version 2 of the experiment applies. The assignment of deduction points to private-account-contributors is possible if the majority of the A-participants of one group (i.e. 2 or 3 A-participants) votes for this option. If only a minority (0 or 1 A-participants) votes for this option, the assignment of deduction points is not possible.

- If the computer determines the vote **not to be considered**, a **random mechanism** determines whether version 1 or version 2 of the experiment applies.

After the vote, all the group members (the three A-participants and the B-participant) will receive information about the vote result and whether it will be considered. Subsequently, all participants learn whether the option to assign deduction points to private-account-contributors will exist in the following 20 periods or not.

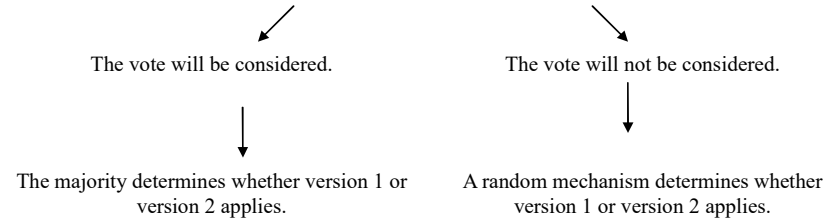
Summary

You will be divided into new groups of 4 (3 A-participants and 1 B-participants). Your fellow group members are participants with whom you have not interacted in part 1. Both group composition and your type remain unchanged for 20 periods.

A-Participants

You decide on whether you wish to proceed with version 1 or version 2 of the experiment.

The computer randomly determines whether the vote will be considered.



In both versions, you decide whether you allocate your point endowment to the private account or to the group account in each of the 20 periods. After every period, you receive information about your point income from both the private account and the group account. If deduction points can be assigned to private-account-contributors, and if you have allocated your point endowment to the private account in the respective period, you will further receive information about whether a B-participant assigned deduction points to you and, if yes, how many.

B-Participants

You receive information about the vote result in your group. You learn whether the vote result will be considered (in which case the majority determines whether version 1 or version 2 applies) or whether it will not be considered (in which case a random mechanism determines whether version 1 or version 2 applies).

- If version 2 applies, you decide on the assignment of deduction points to private-account-contributors prior to the beginning of the 20 periods. The decision on the deduction points applies according to the actual number of private-account-contributors in each of the 20 periods. Afterwards, you are asked to complete a questionnaire.
- If version 1 applies, you are asked to complete a questionnaire.

In both versions, after the A-participants have made their decisions in the 20 periods, you receive information about the A-participants' point allocation, the sum of the assigned deduction points to the private-account-contributors and about your own total point income in each of the periods.

At the end of the experiment, you will receive all payoff-relevant information from part 1 and 2 of the experiment. We kindly ask you to remain seated until you are called.

In this appendix we present a translation of the original German instructions.

Belief Elicitation

These instructions were presented on the participant's screen.

B-type: Belief distribution on A-types' contribution behavior

The A-subjects are now deciding how to use their endowment points in each of the 20 periods. At the same time we would like to ask you to indicate your belief about how often the cases 1-4 will occur in these 20 periods. At the end of the experiment you will receive 10 points for each correct belief. If e.g., your belief about how often case 1 occurs is in line with the actual occurrence in the 20 periods, then you receive 10 points.

A-type: Belief distribution on B-types' punishment behavior

The B-subjects are now deciding on the deduction points. At the same time we would like to ask you to indicate your belief about how many deduction points the B-type assigns in the respective cases. Please indicate for each of the three fields in the table what you think the B-type entered. For each correct belief, i.e., if your belief is in line with the actual entry of the B-type in the respective field, you will receive 10 points.

Questionnaire

These questions were presented on the participant's screen.

B-type: if modified PGG is implemented

1 - Please indicate on a scale from 0 to 10 to what extent you agree with the following statements. 0=I do not agree at all, 10=I completely agree...
If I assign deduction points...
...I feel desired in my role by the A-subjects.
...I feel obligated towards the A-subjects in my role.
...I feel comfortable in my role.

2 - If I make decision that affects other people, it is important for me that they are fine with me having this decision power. [agree/disagree]

3 - Imagine the case that in a period 2 A-subjects chooses the group account

and 1 A-subject chooses the private account. The B-subjects decides to assign one deduction point to the private account contributor so that his income is reduced by 3 points. How likely do you think it is that the private account contributor will choose the group account in the next period [0%, 10%, 20%, ... ,100%]?

4 - The decision of the A-subject who chose the private account is not fair [agree/disagree].

5 - I assigned deduction points in order to...
...change the behavior of the respective A-subject.
...signal that I dispraise the choice of the private account.
...reduce income differences between A-subjects.

B-type: if base PGG is implemented

1 - Please indicate on a scale from 0 to 10 to what extent you agree with the following statement. 0=I don't agree at all, 10=I completely agree...
Consider the case in which two A-subjects choose the group account and one A-subject chooses the private account in a given period. The B-subject decides to assign a deduction points to the private account contributor so that his income is reduced by 3 points.

2 - The decision of the A-subject that chose the private account is not fair [agree/disagree].

A-type: independent of which game is implemented

1 - Please explain in detail how you made your voting decision over version 1 and version 2.

2 - Under which conditions would you rather have voted for *version 2 (with deduction rule) (version 1)*? If the maximum amount of deduction points (in the experiment max 9 points)...
...would have been higher: more than 9 points.
...would have been lower: less than 9 points.
...didn't play a role for my decision.

3 - If the consequence of a deduction points for the income of the A-subject...
...would have been higher: 1 deduction point would have reduced the income by more than 3 points.
...would have been lower: 1 deduction point would have reduced the income by less than 3 points.
...didn't play a role for my decision.

A- and B-type: general questions

1 - Please indicate your gender.
2 - Please indicate your age.
3 - Please indicate your field of study.
4 - In how many experiment have you already participated?
5 - Please describe in detail to what extent you found the instructions and the experiment comprehensible. Was something difficult to understand or not clear? If yes, what was this?

Chapter 5

Conclusion

This dissertation has examined political institutions, their potential to aggregate information in the presence of heterogeneous preferences, and their capability to create legitimacy by providing participation in democratic procedures.

Two of the three chapters have focused on studying behavior within a given institution (committee decision-making in Chapter 2 and social dilemma with external sanctions in Chapter 4). In contrast, Chapter 3 has studied one specific behavioral motive, social image concerns for sharing truthfully endogenous information in a simple setting. In this conclusion I present the main findings of each chapter and discuss the results in a wider context. Given the two focuses I will first discuss Chapter 2 and 4 before I turn to Chapter 3.

In Chapter 2, I analyzed the potential of communication to aggregate information in a committee featuring known heterogeneous preference types, prevote communication and majority rule. I studied whether social preferences or cognitive constrains drive the (non-)existence of strategic communication. The results revealed that aggregate behavior is not consistent with any of the models assuming homogenous agents (social preferences or naïve voting). Further disaggregating results, the data supported a model of cognitive heterogeneity assuming two cognitive sophistication levels. The numerically dominant naive subjects tell the truth and use a decision heuristic, i.e., vote in line with the majority of signals. In contrast, sophisticated subjects follow their type-specific payoff-maximizing decision rule and lie in a way that allows them to influence the committee's decision in their favor.

What are the implications of the observed behavior for the economic analysis of committees? We know that specific features of the design, e.g., the voting rule, existence of pre-vote communication, simultaneous vs. sequential votes, often have a crucial impact on the equilibrium strategies in the Condorcet Jury Game (Palfrey, 2016). However, the equilibrium calculus in the basic game and its variations can be quite intimidating, featuring the application of Bayes' rule, pivotal voter calculus and the need to factor in the strategies of other voters. While there is abundant evidence that many individuals often deviate from standard equilibrium play, e.g., they update beliefs in a way that violate Bayes' rule (e.g., Kahneman and Tversky, 1973; Slovic and Lichtenstein, 1971), have difficulty to condition their decision on hypothetical events (Esponda and Vespa, 2014), and exhibit social preferences (e.g., Engelmann and Strobel, 2004b; Fehr and Schmidt, 1999), we do not know how these factors interact with the institutions governing committee-decision making. Consequently, specific research on committees is needed to learn whether individuals react sensitively to institutional changes.

Research on committee decision-making has demonstrated that aggregate behavior represents features of Nash equilibrium as well as naïve play (Palfrey, 2016). Guarnaschelli, McKelvey, and Palfrey (2000) and Goeree and Yariv (2011) find evidence for strategic voting (with a noise component), but also diversity in individual choices (Guarnaschelli, McKelvey, and Palfrey, 2000). The results discussed in Chapter 2 also display heterogeneity. More importantly, I have found little evidence of strategic communication. This indicates that while subjects may comprehend strategic voting, many are not able to do another step of backward induction that is necessary to understand the incentives of strategic communication, i.e., to find the optimal communication strategy given the voting strategy. Furthermore, learning is difficult. Given that many individuals do not perfectly play strategic voting in the experiment, the incentives for strategic communication are blurred.

Decision making in these environments is highly complex. Hence, in Chapter 2 I studied the different steps of decision-making to identify at which stage deviations from standard equilibrium play occur. Thereby, it could be tested whether the group composition (homogenous vs. heterogeneous) affected individuals' voting decision as suggested by previous research (Goeree and Yariv, 2011). I found no evidence in support of this claim. To understand individual behavior in complex environments, the results stress the need to break up a game into its different components, elicit beliefs and conduct individual comprehensive tasks. By identifying key problems (e.g. the difficulty of finding the

optimal decision-rule), we can pin down institutional challenges. For instance, Chapter 2 shows that further research should explore mechanisms that could prevent the exploitation of naïve individuals.

Chapter 4 has also studied behavior within an institution and used a public goods game with third-party punishment as a metaphor for social norm compliance (Ledyard, 1995). I tackled the research question of how the institutional process that implements third-party punishment affects sanctioning and, in turn, compliance behavior. To this end, I compared the behavior of third-party punishers who are elected by the group the punisher is supposed to sanction to that of third-party punishers who are appointed by chance. The results showed that institutional endogeneity leads to milder sanctions and a higher effectiveness of sanctions.

These findings contribute to previous research in various ways. First, recent research has shown that the effectiveness of fixed sanctions in promoting socially-desirable behavior depends not only on its efficiency but also on the extent to which the sanctioning authority is perceived as just or fair (Dal Bó, Foster, and Putterman, 2010; Tyran and Feld, 2006). The results of my study demonstrated that the sanctioning authority anticipated the impact of legitimacy on effectiveness, therefore reducing her sanctions. This had important implications. While the endogenous sanctions have initially lead to more cooperation, the exogenous and the endogenous environments exhibited identical outcomes, both in terms of cooperation and efficiency, overall. These results suggest that the third-party may have reduced her sanctions too much. Given that third-party punishers had no experience in the task, further research could explore how cooperation and efficiency evolve over time when the punisher has the opportunity to adjust her sanctions.

The results also demonstrate the need to integrate legitimacy into theoretical models to develop a full equilibrium analysis. This would allow to theoretically derive how addressees and addressers of sanctions respond to changes in legitimacy. In Chapter 4, I have provided a starting point for how legitimacy could be included in a utility function. The effectiveness of punishment is assumed to be magnified or attenuated by the legitimacy of the punisher who assigns the sanction.

Closely related are two further open questions. The first question posits which channels other than a vote could legitimize an institution. So far existing research has identified that the procedure of implementing an institution (Baldassarri and Grossman, 2011; Dal Bó, Foster, and Putterman, 2010; Tyran

and Feld, 2006), the method by which enforcers are compensated (Dickson, Gordon, and Huber, 2015), and the degree of transparency (Dickson, Gordon, and Huber, 2015) can constitute institutional sources of legitimacy. Developing an overarching account of legitimacy is however an open avenue for further research.

Secondly, the chapter triggers the question of how legitimacy should be measured. In social psychology, political science, and law, survey measures of attitudes about institutional fairness or support for an institution are commonly used (Caldeira and Gibson, 1992; Tyler, 2006). In my study, I used instead a behavioral measure. The anticipated effectiveness of punishment is taken as a proxy for the perceived legitimacy from the perspective of the third-party. This approach captures the notion from Tyler (2006) that legitimacy can enhance the effectiveness of institutions. Dickson, Gordon, and Huber (2015) propose an alternative measure – the costly decisions of public good players in assisting or hindering third-party sanctions. Their approach represents the idea that legitimacy entails obligation or duty to the authority (Hart, 1994). Both behavioral measures include monetary incentives, so the problem of giving desirable answers that survey measures face is reduced. The obvious next step would be to analyze how these different behavioral and survey measures correlate and eventually develop an experimentally validated survey tool to measure legitimacy.

While in the two latter chapters I examined how individuals react to procedures and rules, in Chapter 3 I studied a behavioral motive. More specifically, I expanded the traditional notion of information in economics by allowing for endogenous information, i.e., the information's precision depends on the sender's ability to extract the true state out of the given information. I studied whether social image considerations originating in the endogenous nature of information may motivate truthful communication in a very simple setting. I have shown in a sender-receiver game with misaligned interests that senders endowed with endogenous information of high social status tell the truth significantly more often compared to senders in a treatment with low social status information. When senders receive exogenous information, the treatment effect between knowledge areas disappears. Thus, the relevant driving factor is the endogeneity of information which gives senders the possibility to signal their expertise to the receiver.

This chapter contributes to two streams of literature, i.e., information economics more generally and social image considerations more specifically. With

regards to the former, I demonstrated that the source of information matters considerably for truthful communication. When information is endogenous, a message reveals personal characteristics and interests of the sender. Why does this person possess this piece of information? Which of her characteristics explain her interests for the subject? Asking this type of questions is natural for most of us in personal relations when getting to know other people, but has been largely neglected by economic theory. It would be important to delineate in how far endogenous information is different from exogenous information. For instance, when information is uncertain, endogenous information may be more likely to give rise to overconfidence than exogenous information.

The reason why the source of information matters are social image concerns. So far the literature has mainly focused on their role for the performance of prosocial activities and the provision of public goods (Bénabou and Tirole, 2006; Harbaugh, 1998; Polborn, 2008). More recently, social image concerns were shown to matter for the desire to signal ability (Ewers and Zimmermann, 2015). This chapter has provided evidence that the weight of social image utility depends on the social status of knowledge transmitted in communication, or put more generally, on the social status of the particular ability. This view follows the concept of identity utility (Akerlof and Kranton, 2000), where social categories are attributed a social status. The social status of expertise can be derived from how the knowledge is evaluated in the social system.

There is one follow-up question that I deem most central for future research. Does a knowledge area exhibiting a negative social status induce lying in situations where the sender has an incentive for truthful communication? A priori, the effect of social status may not be symmetric as social image considerations may interact with lying aversion. For instance, in a situation where incentives for lying are present and information exhibits a positive social status, an individual may face some lying aversion which is independent of the information to be transmitted. The combined effect of lying costs and positive social image may be large enough to deter individuals from lying. On the contrary, in situations with truth-telling incentives and negative social status of information, the lying aversion works against the social image effect. Thus, the negative social status may have a smaller or even no effect on truth-telling behavior. Further empirical research is needed to answer this question.

Bibliography

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond (2016). "Preferences for truth-telling". In: *IZA Discussion Paper* 10188.

Ai, Chunrong and Edward C Norton (2003). "Interaction terms in logit and probit models". In: *Economics letters* 80.1, pp. 123–129.

Akerlof, George A. and Rachel E. Kranton (2000). "Economics and Identity". In: *The Quarterly Journal of Economics* 115.3, pp. 715–753.

Almenberg, Johan, Anna Dreber, Coren Apicella, and David G Rand (2010). "Third party reward and punishment: group size, efficiency and public goods". In: *Psychology of Punishment, Nova Publishing*.

Andreoni, James, William Harbaugh, and Lise Vesterlund (2003). "The Carrot or the Stick: Rewards, Punishments, and Cooperation". In: *The American Economic Review* 93.3, pp. 893–902.

Anwar, Shamena, Patrick Bayer, and Randi Hjalmarsson (2015). "Politics in the Courtroom: Political Ideology and Jury Decision Making". In:

Argenziano, Rossella, Sergei Severinov, and Francesco Squintani (2016). "Strategic Information Acquisition and Transmission". In: *American Economic Journal: Microeconomics* 8.3, pp. 119–55.

Austen-Smith, David (1993). "Interested experts and policy advice: Multiple referrals under open rule". In: *Games and Economic Behavior* 5.1, pp. 3–43.

— (1994). "Strategic Transmission of Costly Information". In: *Econometrica* 62.4, pp. 955–963.

Austen-Smith, David and Jeffrey S Banks (1996). "Information aggregation, rationality, and the Condorcet jury theorem". In: *The American Political Science Review* 90.1, pp. 34–45.

Austen-Smith, David and Timothy J Feddersen (2006). "Deliberation, preference uncertainty, and voting rules". In: *The American Political Science Review* 100.2, pp. 209–217.

Baldassarri, Delia and Guy Grossman (2011). "Centralized sanctioning and legitimate authority promote cooperation in humans". In: *Proceedings of the National Academy of Sciences* 108.27, pp. 11023–11027.

Battaglini, Marco (2002). "Multiple referrals and multidimensional cheap talk". In: *Econometrica* 70.4, pp. 1379–1401.

Battaglini, Marco, Rebecca B Morton, and Thomas R Palfrey (2008). "Information aggregation and strategic abstention in large laboratory elections". In: *The American Economic Review* 98.2, pp. 194–200.

— (2010). "The swing voter's curse in the laboratory". In: *The Review of Economic Studies* 77.1, pp. 61–89.

Becker, Gary S. (1968). "Crime and Punishment: An Economic Approach". In: *Journal of Political Economy* 76, pp. 169–217.

Bénabou, Roland and Jean Tirole (2006). "Incentives and Prosocial Behavior". In: *American Economic Review* 96.5, pp. 1652–1678.

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch (2014). "hroot: Hamburg registration and organization online tool". In: *European Economic Review* 71, pp. 117–120.

Bolton, Gary E. and Axel Ockenfels (2000). "ERC: A Theory of Equity, Reciprocity, and Competition". In: *The American Economic Review* 90.1, pp. 166–193.

Botelho, Anabela, Glenn W Harrison, Lígia Pinto, and Elisabet E Rutström (2005). "Social norms and social choice". Unpublished.

Brandts, Jordi and Gary Charness (2011). "The strategy versus the direct-response method: a first survey of experimental comparisons". In: *Experimental Economics* 14.3, pp. 375–398.

Burks, Stephen V., Jeffrey P. Carpenter, Lorenz Goette, and Aldo Rustichini (2013). "Overconfidence and Social Signalling". In: *The Review of Economic Studies*.

Cai, Hongbin and Joseph Tao-Yi Wang (2006). "Overcommunication in strategic information transmission games". In: *Games and Economic Behavior* 56.1, pp. 7–36.

Caldeira, Gregory A. and James L. Gibson (1992). "The etiology of public support for the Supreme Court". In: *American journal of political science*, pp. 635–664.

Camerer, Colin F. (2003). *Behavioral Game Theory - Experiments in Strategic Interaction*. Princeton University Press.

Charness, Gary, Ramón Cobo-Reyes, and Natalia Jiménez (2008). "An investment game with third-party intervention". In: *Journal of Economic Behavior & Organization* 68.1, pp. 18–28.

Charness, Gary and Matthew Rabin (2002). "Understanding social preferences with simple tests". In: *The Quarterly Journal of Economics* 117.3, pp. 817–869.

Chaudhuri, Ananish (2011). "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature". In: *Experimental Economics* 14.1, pp. 47–83.

Chen, Josie I (2014). "Obedience to Rules with Mild Sanctions: The Roles of Peer Punishment and Voting". Unpublished.

Clippel, Geoffroy de (2014). "Behavioral implementation". In: *The American Economic Review* 104.10, pp. 2975–3002.

Condorcet, Nicolas (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie royale.

Coughlan, Peter J (2000). "In defense of unanimous jury verdicts: Mistrials, communication, and strategic voting". In: *The American Political Science Review* 94.02, pp. 375–393.

Crawford, Vincent P and Nagore Iriberri (2007). "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" In: *Econometrica* 75.6, pp. 1721–1770.

Crawford, Vincent P and Joel Sobel (1982). "Strategic information transmission". In: *Econometrica: Journal of the Econometric Society*, pp. 1431–1451.

Dal Bó, P, A Foster, and L Putterman (2010). "Institutions and behavior: Experimental evidence on the effects of democracy". In: *The American Economic Review* 100.5, pp. 2205–2229.

Daruvala, Dinky (2010). "Would the right social preference model please stand up!" In: *Journal of Economic Behavior & Organization* 73.2, pp. 199 –208.

Deimen, Inga, Felix Ketelaar, and Mark T Le Quement (2015). "Consistency and communication in committees". In: *Journal of Economic Theory* 160, pp. 24–35.

Devine, Dennis J, Laura D Clayton, Benjamin B Dunford, Rasmy Seying, and Jennifer Pryce (2001). "Jury decision making: 45 years of empirical research on deliberating groups." In: *Psychology, public policy, and law* 7.3, p. 622.

Dickson, Eric S, Sanford C Gordon, and Gregory A Huber (2009). "Enforcement and compliance in an uncertain world: An experimental investigation". In: *The Journal of Politics* 71.04, pp. 1357–1378.

— (2015). "Institutional Sources of Legitimate Authority: An Experimental Investigation". In: *American Journal of Political Science* 59.1, pp. 109–127.

Dickson, Eric S, Catherine Hafer, and Dimitri Landa (2008). "Cognition and strategy: a deliberation experiment". In: *The Journal of Politics* 70.04, pp. 974–989.

Diekmann, Andreas, Wojtek Przepiorka, and Heiko Rauhut (2015). "Lifting the veil of ignorance: An experiment on the contagiousness of norm violations". In: *Rationality and Society* 27.3, pp. 309–333.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner (2011). "Individual risk attitudes: Measurement, determinants, and behavioral consequences". In: *Journal of the European Economic Association* 9.3, pp. 522–550.

Ellingsen, Tore and Robert Östling (2010). "When does communication improve coordination?" In: *The American Economic Review* 100.4, pp. 1695–1724.

Engel, Christoph (2014). "Social preferences can make imperfect sanctions work: Evidence from a public good experiment". In: *Journal of Economic Behavior & Organization* 108, pp. 343–353.

Engel, Christoph and Lilia Zhurakhovska (2013). "Words substitute fists: Justifying punishment in a public good experiment". In: *Preprints of the Max Planck Institute for Research on Collective Goods* 2013/06.

Engelmann, Dirk and Martin Strobel (2004a). "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments". In: *American economic review*, pp. 857–869.

— (2004b). "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments". In: *The American Economic Review*, pp. 857–869.

Ertan, Arhan, Talbot Page, and Louis Putterman (2009). "Who to punish? Individual decisions and majority rule in mitigating the free rider problem". In: *European Economic Review* 53.5, pp. 495–511.

Esponda, Ignacio and Emanuel Vespa (2014). "Hypothetical thinking and information extraction in the laboratory". In: *American Economic Journal: Microeconomics* 6.4, pp. 180–202.

Ewers, Mara and Florian Zimmermann (2015). "Image and Misreporting". In: *Journal of the European Economic Association* 13.2, pp. 363–380.

Feddersen, Timothy and Wolfgang Pesendorfer (1998). "Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting". In: *The American Political Science Review* 92.01, pp. 23–35.

Feddersen, Timothy J and Wolfgang Pesendorfer (1996). "The swing voter's curse". In: *The American Economic Review*, pp. 408–424.

Fehr, Ernst and Urs Fischbacher (2004a). "Social norms and human cooperation". In: *Trends in cognitive sciences* 8.4, pp. 185–190.

— (2004b). "Third-party punishment and social norms". In: *Evolution and human behavior* 25.2, pp. 63–87.

Fehr, Ernst and Simon Gächter (2000). "Cooperation and punishment in public goods experiments". In: *The American Economic Review* 90.4, pp. 980–994.

Fehr, Ernst and Klaus M. Schmidt (1999). "A Theory of Fairness, Competition and Cooperation". In: *The Quarterly Journal of Economics* 114.3, pp. 817–868.

Feld, Lars P and Bruno S Frey (2002). "Trust breeds trust: How taxpayers are treated". In: *Economics of Governance* 3.2, pp. 87–99.

Fischbacher, Urs (2007). "z-Tree: Zurich toolbox for ready-made economic experiments". In: *Experimental Economics* 10.2, pp. 171–178.

Fischbacher, Urs and Franziska Föllmi-Heusi (2013). "Lies in disguise – an experimental study on cheating". In: *Journal of the European Economic Association* 11.3, pp. 525–547.

Fischbacher, Urs and Simon Gaechter (2010). "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments". In: *American Economic Review* 100.1, pp. 541–56.

Fischbacher, Urs, Arye Hillman, and Heinrich Ursprung (2015). "Introduction to Special Issue 'Behavioral Political Economy'". In: *European Journal of Political Economy* 40, Part B, p. 207.

Ford, P. Dianne and Sandy D Staples (2006). "Perceived value of knowledge: the potential informer's perception". In: *Knowledge Management Research & Practice* 4.1, pp. 3–16.

Furnham, Adrian and Tomas Chamorro-Premuzic (2006). "Personality, intelligence and general knowledge". In: *Learning and Individual Differences* 16.1, pp. 79–90.

Gerardi, Dino (2000). "Jury verdicts and preference diversity". In: *The American Political Science Review* 94.02, pp. 395–406.

Gerardi, Dino and Leeat Yariv (2007). "Deliberative voting". In: *Journal of Economic Theory* 134.1, pp. 317–338.

Gettier, Edmund L. (1963). "Is Justified True Belief Knowledge?" In: *Analysis* 23.6, pp. 121–123.

Gneezy, Uri (2005). "Deception: The role of consequences". In: *American Economic Review*, pp. 384–394.

Gneezy, Uri, Bettina Rockenbach, and Marta Serra-Garcia (2013). "Measuring lying aversion". In: *Journal of Economic Behavior & Organization* 93.0, pp. 293–300.

Goeree, Jacob K and Charles A Holt (2004). "A model of noisy introspection". In: *Games and Economic Behavior* 46.2, pp. 365–382.

Goeree, Jacob K and Leeat Yariv (2011). "An experimental study of collective deliberation". In: *Econometrica* 79.3, pp. 893–921.

Grosser, Jens and Michael Seebauer (2016). "The curse of uninformed voting: An experimental study". In: *Games and Economic Behavior* 97, pp. 205–226.

Grossman, Guy and Delia Baldassarri (2012). "The Impact of Elections on Cooperation: Evidence from a Lab-in-the-Field Experiment in Uganda". In: *American Journal of Political Science* 56.4, pp. 964–985.

Guarnaschelli, Serena, Richard D McKelvey, and Thomas R Palfrey (2000). "An experimental study of jury decision rules". In: *The American Political Science Review* 94.02, pp. 407–423.

Gunnthorsdottir, Anna, Daniel Houser, and Kevin McCabe (2007). "Disposition, history and contributions in public goods experiments". In: *Journal of Economic Behavior & Organization* 62.2, pp. 304–315.

Gürerk, Özgür, Bernd Irlenbusch, and Bettina Rockenbach (2006). "The competitive advantage of sanctioning institutions". In: *Science* 312.5770, pp. 108–111.

— (2009). "Motivating teammates: The leader's choice between positive and negative incentives". In: *Journal of Economic Psychology* 30.4, pp. 591–607.

Hafer, Catherine and Dimitri Landa (2007). "Deliberation as self-discovery and institutions for political speech". In: *Journal of Theoretical Politics* 19.3, pp. 329–360.

Hanssen, F. Andrew (1999). "The Effect of Judicial Institutions on Uncertainty and the Rate of Litigation: The Election Versus Appointment of State Judges". In: *The Journal of Legal Studies* 28.1, pp. 205–232.

Harbaugh, William T. (1998). "The prestige motive for making charitable transfers". In: *The American Economic Review* 88.2, pp. 277–282.

Harsanyi, John C and Reinhard Selten (1988). "A general theory of equilibrium selection in games". In: *MIT Press Books* 1.

Hart, Herbert Lionel Adolphus (1994). *The concept of law*. 2nd ed. Oxford: Clarendon Press.

Henrich, Joseph et al. (2006). "Costly punishment across human societies". In: *Science* 312.5781, pp. 1767–1770.

Jackson, John D and Nikolay P Kovalev (2006). "Lay adjudication and human rights in Europe". In: *Colum. J. Eur. L.* 13, p. 83.

Kahneman, Daniel and Amos Tversky (1973). "On the psychology of prediction." In: *Psychological review* 80.4, p. 237.

Kamei, Kenju (2016). "Democracy and resilient pro-social behavioral change: an experimental study". In: *Social Choice and Welfare* 47.2, pp. 359–378.

Kamei, Kenju, Louis Putterman, and Jean-Robert Tyran (2015). "State or nature? Endogenous formal versus informal sanctions in the voluntary provision of public goods". In: *Experimental Economics* 18.1, pp. 38–65.

Kartik, Navin (2009). "Strategic communication with lying costs". In: *The Review of Economic Studies* 76.4, pp. 1359–1395.

Kartik, Navin, Marco Ottaviani, and Francesco Squintani (2007). "Credulity, lies, and costly talk". In: *Journal of Economic theory* 134.1, pp. 93–116.

Kawagoe, Toshiji and Hirokazu Takizawa (2012). "Level-k analysis of experimental centipede games". In: *Journal Of Economic Behavior & Organization* 82.2, pp. 548–566.

Krupka, Erin L and Roberto A Weber (2013). "Identifying social norms using coordination games: Why does dictator game sharing vary?" In: *Journal of the European Economic Association* 11.3, pp. 495–524.

Kurzban, Robert, Peter DeScioli, and Erin O'Brien (2007). "Audience effects on moralistic punishment". In: *Evolution and Human behavior* 28.2, pp. 75–84.

Le Quement, Mark (2013). "Communication compatible voting rules". In: *Theory and decision* 74.4, pp. 479–507.

Le Quement, Mark and Venuga Yokeeswaran (2015). "Subgroup deliberation and voting". In: *Social Choice and Welfare*, pp. 1–32.

Ledyard, J. (1995). "Public goods: A survey of experimental research". In: *The Handbook of Experimental Economics*. Ed. by John H. Kagel and Alvin E. Roth. Vol. 111, p. 194.

Lergetporer, Philipp, Silvia Angerer, Daniela Glätzle-Rützler, and Matthias Sutter (2014). "Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation". In: *Proceedings of the National Academy of Sciences* 111.19, pp. 6916–6921.

MacCoun, Robert J (1989). "Experimental research on jury decision-making". In: *Science* 244.4908, pp. 1046–1050.

Machery, Edouard et al. (2015). "Gettier Across Cultures". In: *Noûs*, pp. 1–20.

Markussen, Thomas, Louis Putterman, and Jean-Robert Tyran (2014). "Self-organization for collective action: An experimental study of voting on sanction regimes". In: *The Review of Economic Studies* 81.1, pp. 301–324.

Martinelli, César (2006). "Would rational voters acquire costly information?" In: *Journal of Economic Theory* 129.1, pp. 225–251.

Mazar, Nina, On Amir, and Dan Ariely (2008). "The dishonesty of honest people: A theory of self-concept maintenance". In: *Journal of marketing research* 45.6, pp. 633–644.

McKelvey, Richard D and Thomas R Palfrey (1995). "Quantal response equilibria for normal form games". In: *Games and Economic Behavior* 10.1, pp. 6–38.

— (1998). "Quantal response equilibria for extensive form games". In: *Experimental economics* 1.1, pp. 9–41.

Meirowitz, Adam (2007). "In defense of exclusionary deliberation: communication and voting with private beliefs and values". In: *Journal of Theoretical Politics* 19.3, pp. 301–327.

Murphy, Ryan O, Kurt A Ackermann, and Michel Handgraaf (2011). "Measuring social value orientation". In: *Judgment and Decision Making* 6.8, pp. 771–781.

Nagel, Rosemarie (1995). "Unraveling in guessing games: An experimental study". In: *The American Economic Review* 85.5, pp. 1313–1326.

Nagin, Daniel S. (1998). "Criminal deterrence research at the outset of the twenty-first century". In: *Crime and justice*, pp. 1–42.

Nicklisch, Andreas, Kristoffel Grechenig, and Christian Thöni (2016). "Information-sensitive Leviathans". In: *Journal of Public Economics* 144, pp. 1–13.

Nikiforakis, Nikos and Helen Mitchell (2014). "Mixing the carrots with the sticks: third party punishment and reward". In: *Experimental Economics* 17.1, pp. 1–23.

Norton, Edward C, Hua Wang, Chunrong Ai, et al. (2004). "Computing interaction effects and standard errors in logit and probit models". In: *Stata Journal* 4, pp. 154–167.

Ostrom, Elinor, James Walker, and Roy Gardner (1992). "Covenants with and without a Sword: Self-governance Is Possible." In: *American Political Science Review* 86.02, pp. 404–417.

Palfrey, Thomas R. (2016). "Experiments in political economy". In: *Handbook of Experimental Economics*. Ed. by J. H. Kagel and A. E. Roth. Vol. 2. Princeton University Press. Chap. Experiments in political economy, pp. 347–434.

Pei, Harry Di (2015). "Communication with endogenous information acquisition". In: *Journal of Economic Theory* 160, pp. 132 –149.

Pennington, Nancy and Reid Hastie (1990). "Practical implications of psychological research on juror and jury decision making". In: *Personality and Social Psychology Bulletin* 16.1, pp. 90–105.

Persico, Nicola (2004). "Committee design with endogenous information". In: *The Review of Economic Studies* 71.1, pp. 165–191.

Piketty, Thomas (1999). "The information-aggregation approach to political institutions". In: *European Economic Review* 43.4, pp. 791–800.

Polborn Mattias, K. (2008). *Competing for Recognition through Public Good Provision.*

Rabin, Matthew (1993). "Incorperating fairness into game theory and economics". In: *The American Economic Review* 83.5, pp. 1281–1302.

Rasmusen, Eric, Manu Raghav, and Mark Ramseyer (2009). "Convictions versus conviction rates: the prosecutor's choice". In: *American Law and Economics Review* 11.1, pp. 47–78.

Rauhut, Heiko (2013). "Beliefs about lying and spreading of dishonesty: Undetected lies and their constructive and destructive social dynamics in dice experiments". In: *PloS one* 8.11, e77878.

Rousseau, Jean-Jacques (1896). *Du contrat social.* Alcan.

Rowley, Charles (1993). *Public choice theory.* Edward Elgar Publishing.

Sánchez-Pagés, Santiago and Marc Vorsatz (2007). "An experimental study of truth-telling in a sender–receiver game". In: *Games and Economic Behavior* 61.1, pp. 86–112.

— (2009). "Enjoy the silence: an experiment on truth-telling". In: *Experimental Economics* 12.2, pp. 220–241.

Saran, Rene (2016). "Bounded depths of rationality and implementation with complete information". In: *Journal of Economic Theory* 165, pp. 517–564.

Schnellenbach, Jan and Christian Schubert (2015). "Behavioral political economy: A survey". In: *European Journal of Political Economy* 40, Part B, pp. 395–417.

Schweizer, Mark (2016). "The civil standard of proof – what is it, actually?" In: *The International Journal of Evidence & Proof* 20.3, pp. 217–234.

Slovic, Paul and Sarah Lichtenstein (1971). "Comparison of Bayesian and regression approaches to the study of information processing in judgment". In: *Organizational behavior and human performance* 6.6, pp. 649–744.

Sommers, Samuel R and Phoebe C Ellsworth (2003). "How Much Do We Really Know about Race and Juries-A Review of Social Science Theory and Research". In: *Chi.-Kent L. Rev.* 78, p. 997.

Stahl, Dale O and Paul W Wilson (1994). "Experimental evidence on players' models of other players". In: *Journal of Economic Behavior & Organization* 25.3, pp. 309–327.

— (1995). "On players models of other players: Theory and experimental evidence". In: *Games and Economic Behavior* 10.1, pp. 218–254.

Stutzer, A (1999). *Demokratieindizes für die Kantone der Schweiz. Institut für Empirische Wirtschaftsforschung, University of Zurich*. Tech. rep. IEW Working paper.

Sutter, Matthias (2009). "Deception Through Telling the Truth?! Experimental Evidence From Individuals and Teams*". In: *The Economic Journal* 119.534, pp. 47–60.

Sutter, Matthias, Stefan Haigner, and Martin G Kocher (2010). "Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations". In: *The Review of Economic Studies* 77.4, pp. 1540–1566.

Tan, Fangfang and Erte Xiao (2014). "Third-Party Punishment: Retribution or Deterrence?" In: *Working Paper of the Max Planck Institute for Tax Law and Public Finance*.

Tocqueville, Alexis Henri C. M. Clérel (1836). *De la démocratie en Amérique*. Paris: Pagnerre.

Tyler, Tom R. (2006). "Psychological perspectives on legitimacy and legitimation". In: *Annu. Rev. Psychol.* 57, pp. 375–400.

Tyran, Jean-Robert and Lars P Feld (2006). "Achieving Compliance when Legal Sanctions are Non-deterrent". In: *The Scandinavian Journal of Economics* 108.1, pp. 135–156.

Utikal, Verena and Urs Fischbacher (2013). "Disadvantageous lies in individual decisions". In: *Journal of Economic Behavior & Organization* 85, pp. 108–111.

Van Weelden, Richard (2008). "Deliberation rules and voting". In: *The Quarterly Journal of Political Science* 3.1, pp. 83–88.

Vanberg, Christoph (2016). "Who never tells a lie?" In: *Experimental Economics*, pp. 1–12.

Wang, Joseph Tao-yi, Michael Spezio, and Colin Camerer (2009). "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth-Telling and Deception in Sender-Receiver Game". In: *American Economic Review, Forthcoming*.

Yamagishi, Toshio (1986). "The provision of a sanctioning system as a public good." In: *Journal of Personality and social Psychology* 51.1, p. 110.

Isabel Marcin

Contact Information	University of Heidelberg Bergheimer Str. 58 69115 Heidelberg, Germany	*Phone:* +49 163 667 5253 *Email:* isabel.marcin@awi.uni-heidelberg.de

Personal Details	Born: June 26, 1986, Würzburg Nationality: German
Current Position	University of Heidelberg, Research Fellow (since 10/2016)
Previous Position	Max Planck Institute for Research on Collective Goods, Bonn, Research Fellow (03/2013-09/2016)
Education	International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World, Ph.D. Candidate in Economics (since 03/2013)
	University of Hamburg, Germany, M.Sc. Politics, Economics and Philosophy (2013)
	University Paris 1 Panthéon-Sorbonne, France, Maîtrise Economics (2009)
	University of Hamburg, Germany, B.Sc. Economics (2009)
Research Visit	New York University, Center for Experimental Social Sciences (09/2015-12/2015) Host: Rebecca Morton
Grants and Awards	PhD scholarship, Max Planck Institute for Research on Collective Goods Bonn (2013-2016)
	M.Sc. Thesis Research grand, University of Hamburg (2012)
	Scholarship of the Carlo Schmid Program for Internships in International Organizations and EU Institutions, German Academic Exchange Service and German National Academic Foundation (09/2009-02/2010)
	Best Student Award for Master Degree (2013) and Bachelor Degree (2009) in Economics at University of Hamburg, Academic Foundation Hamburg
	Scholarship of the German Academic Exchange Service, European Excellency Program in Economics (09/2008-06/2009)
	Scholarship of Evangelisches Studienwerk Villigst (Protestant Academic Merit Foundation) (01/2006-09/2012)

October, 12th 2016 Isabel Marcin

Erklärung nach §4 Abs. 1 PromO:

Hiermit erkläre ich,

1. dass mir die geltende Promotionsordnung bekannt ist;
2. dass ich die Dissertation selbst angefertigt, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung ubernommen und alle von mir benutzten Hilfsmittel, personlichen Mitteilungen und Quellen in meiner Arbeit angegeben habe;
3. dass ich bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskriptes keine unzulässige Hilfe in Anspruch genommen habe;
4. dass ich nicht die Hilfe eines Promotionsberaters in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen;
5. dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prufung eingereicht habe;
6. dass ich nicht die gleiche, eine in wesentlichen Teilen ahnliche oder eine andere Abhandlung bei einer anderen Hochschule bzw. anderen Fakultat als Dissertation eingereicht habe

Isabel Marcin 12.10.2016, Heidelberg